V. Communications Systems Research

TELECOMMUNICATIONS DIVISION

A. Digital Clean-Up Loop Transponder for Sequential Ranging System, *W. J. Hurd*

1. Introduction

The range from a ground tracking station to a spacecraft is determined by measuring the round-trip time required for a ranging code sequence signal to be transmitted from the station to the spacecraft and transponded back to the ground station. In typical existing spacecraft, the ranging transponders demodulate and limit the received ranging signal and then remodulate it onto a different frequency RF carrier for retransmission to the earth. A major drawback of such transponders is that both the received ranging signal and the receiver noise are transmitted on the downlink. This means that when the uplink ranging channel SNR is below unity, most of the downlink ranging power is wasted on transmitting receiver noise rather than ranging signal. The problem will be extremely severe on outer planet missions. For a *Mariner*-type spacecraft at a range of 31 AU, the approximate distance to Neptune, the uplink SNR will be as low as −46 dB, even using a 210-ft transmitting antenna and a 400-kW transmitter.

To overcome the problem of retransmission of receiver noise, a digital signal processing system has been developed to regenerate the ranging signal in the spacecraft for transmission to the ground. The system, called a clean-up loop, phase locks on the received ranging signal and generates a replica of the received ranging code. The regenerated ranging signal has no amplitude noise, but a small amount of phase jitter, corresponding to the phase error in the phase-locked regeneration loop. This phase noise can be averaged out on the ground. Since little downlink transmitter power is consumed by transmission of the phase noise, the increase in downlink SNR due to the clean-up loop is approximately equal to the ratio of uplink receiver noise to signal, or up to 46 dB. Less power need be devoted to the ranging signal, on both the uplink and the downlink, and accurate ranging can be accomplished in a shorter time.

2. Sequential Component Ranging System

The clean-up loop is designed to operate in conjunction with a binary coded sequential code-component ranging system, called the μ-system (SPS 37-52, Vol. II, pp. 46–49;

SPS 37-57, Vol. III, pp. 72–81, and Ref. 1) rather than with the parallel component pseudonoise system. The sequential component system, without a clean-up loop, was used successfully in the post-encounter phases of *Mariners VI* and *VII*. In the μ-system, square waves at successively lower frequencies are transmitted sequentially. First the range is measured to high resolution, but with 2 μsec ambiguities, using a 500-kHz square wave. Then a 250-kHz square wave is transmitted, and half of the ambiguities are eliminated. Lower and lower frequencies are transmitted until all ambiguities are resolved. The clean-up loop transponder operates by phase locking on each code component frequency as it is received by the spacecraft. A square wave in phase with the received signal is generated for modulation onto the downlink carrier and transmission to earth.

The information bearing components of the ranging code, $c_k(t)$, are square waves in phase with sine waves, and are defined by

$$c_k(t) = \text{sgn}\left[\sin\left(2^{-k} 2\pi f_0 t\right)\right] \qquad \text{for } k = 1, 2, \cdots, K$$

where f_0 is approximately 1 MHz. The number of code components, K, is chosen to be large enough so that the period of c_K is sufficiently long to resolve all *a priori* uncertainty in range. In the prototype clean-up loop, K was chosen to be 16.

When using the clean-up loop, c_1 is not transmitted by itself, but is used as a suppressed subcarrier for the other components. Its frequency, 500 kHz, determines the basic resolution of the system. The actual sequence of transmitted waveforms is

$$\{w_k(t), k = 2, 3, \cdots, 16\}$$

where

$$w_k = c_1 c_k \qquad \text{for } k = 2, 3, \cdots, 9, 11, 13, 15$$

and

$$w_k = c_1 c_k c_7 \qquad \text{for } k = 10, 12, 14, 16$$

The purpose of introducing the third component, c_7, into w_k for $k = 10$, 12, 14, and 16 is so that adjacent components decorrelate in a shorter length of time. Components c_8 and c_9 are orthogonal over 256 μsec, the period of c_8, and all other adjacent pairs are orthogonal over 128 μsec or less.

Operationally, the first component (w_2) is transmitted for a time long enough for the clean-up loop to acquire frequency and phase lock on this component. The transmitter then switches to w_3, then w_4, etc. The clean-up loop must determine when the received signal changes from w_k to w_{k+1}, make a binary decision as to the phase of w_{k+1}, and change the phase-locked loop (PLL) reference signal to track the new component. Note that w_{k+1} can arrive in one of two possible phase relationships with respect to w_k, since consecutive transitions of c_k correspond alternately to positive and negative transitions of c_{k+1}. The determination of the phase relationships between the c_k corresponds to resolution of the ambiguities in range measurement.

Since the downlink SNR is normally lower than the uplink SNR, the time required to measure range to the desired accuracy is determined primarily by the downlink SNR, as discussed by Goldstein (SPS 37-52, Vol. II). However, because the decision time parameters of the clean-up loop are determined by the uplink design point SNR, each code component must be transmitted for some minimum length of time, even when the SNR is high. Conservatively, the minimum time is about 2 to 3 min for acquisition of the first component and about 12 seconds for each of the other components, for a total of 5–6 min. At the lowest SNRs, it may be desirable to allow more time for initial acquisition, especially if the frequency offset is not small.

3. Implementation of Clean-up Loop

In this section we describe the first implementation of the clean-up loop. Methods for improving the particular design and possible changes in design philosophy which would result in major changes are presented in *Subsection 6*.

A fundamental block diagram of the clean-up loop, which would probably not change with any modifications, is shown in Fig. 1. The loop has three major portions: the PLL, the lock detector, and the code generator

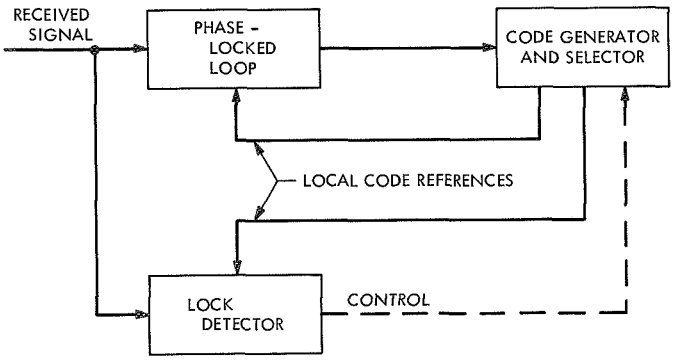

Fig. 1. Clean-up loop simplified block diagram

and selectors. The present implementation is entirely digital, with the exception of the voltage-controlled oscillator (VCO), which is controlled by a digital-to-analog converter (DAC). Because the ranging signal in *Mariner*-type transponders is most readily available at the output of a hard limiter, the input to the clean-up loop is assumed to be binary. This simplifies the implementation of the PLL and the lock detectors, since products can be realized with exclusive or gates and additions and accumulations can be realized with counters. The input signal S is sampled at a 4-MHz rate, synchronously with the local reference clock. Every fourth sample occurs at the transition of $\hat{c}_1$, the locally generated replica of c_1. These samples are used by the phase detector, and the three samples between each transition of $\hat{c}_1$ are used by the lock detector.

a. Phase-locked loop. The PLL consists of the phase detector, a linear digital first order filter, a nonlinear digital low pass filter, a DAC and a VCO, as shown in Fig. 2. Since the nonlinear filter has no effect on the loop dynamics except as a gain control, the loop can be modeled as a perfect integrator second-order loop, with open-loop transfer function

$$F(s) = \frac{A}{s}\left(K_1 + \frac{K_2}{s}\right) \qquad (1)$$

where A, the phase detector gain, depends on the input SNR. The bandwidth of the loop is varied during acquisition by changing the parameters of the digital filters upon command of the lock detector. To maintain a fixed damping ratio, K_1 is varied in proportion to loop bandwidth and K_2 is varied in proportion to the square of loop bandwidth. This is accomplished by varying the gain of the nonlinear filter in direct proportion to loop bandwidth, which changes both K_1 and K_2, and by varying the integrator gain in the linear filter directly with loop bandwidth, which affects only K_2.

Phase detector. The phase detector is of the transition tracking type (Ref. 2). An estimate of the sign of the phase error is made each time a transition occurs in the local replica of the code component being tracked. The code generator and selector informs the phase detector when such a transition occurs and its direction. The product of the transition direction and the sample of the received signal taken at the time of the transition forms an estimate of the sign of the phase error. For example, if the sample is positive (negative) and the transition is negative going, then the sampling time is most likely early (late) compared to the true transition time in the received signal. The phase estimate is implemented simply as the exclusive-or of the input signal sample and the value of $\hat{w}_k$ at the instant immediately preceding the possible transition time. This value is gated into the nonlinear filter only when $\hat{w}_k$ changes, i.e., whenever there truly is a transition.

Nonlinear filter. The nonlinear filter consists of two up/down counters and some gating logic (SPS 37-61, Vol. II, pp. 63–68). The inputs to the filter are all 1, 0, or -1, and the outputs are all 1 or -1; however, the filter low-frequency gain can easily be varied by changing the number of stages in one of the counters. For the wide, medium, and narrow loop bandwidths, respectively, the gain is either 32, 8, or 2. To achieve a gain of 2^N, i.e., an average output equal to 2^N times the average

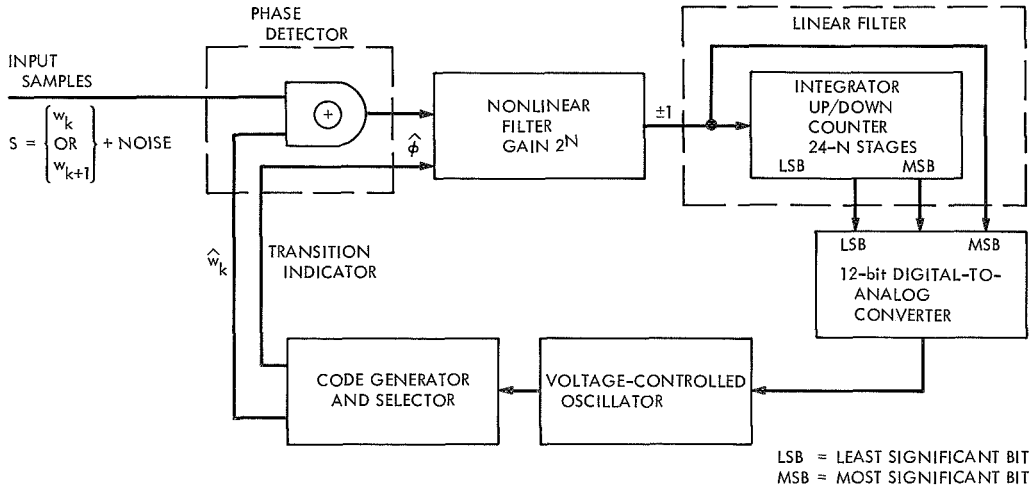

Fig. 2. Implementation of phase-locked loop

input, the number of $+1$ outputs less the number of -1 outputs must be 2^N times the number of $+1$ inputs less the number of -1 inputs. To accomplish this, a count stored in the two counters is incremented or decremented by 2^N for each $+1$ or -1 input, and decremented or incremented by one for each $+1$ or -1 output. The filter output is the sign of the value in the counters.

Linear filter. Since the output of the nonlinear filter is always either $+1$ or -1, the integrator portion of the loop filter can be simply implemented with an up/down counter. The number of stages in the counter is varied with loop bandwidth to control K_2. When the nonlinear filter has gain 2^N, the number of stages in the integrator-counter is $24 - N$, so that K_2 varies as 2^{2N}. The proportional control component of the linear filter is fixed, so K_1 varies only as the nonlinear filter gain, 2^N.

Digital-to-analog converter. The proportional control and integrated components of the loop filter output are summed in one digital-to-analog converter. This can be accomplished because of the careful scaling of the digital filters. This scaling is such that the proportional control component controls the most significant bit of the DAC, and the integrator-counter controls the remaining 11 bits. The actual voltage output of the DAC is of course made compatible with the range and gain of the VCO.

b. Code generator and selector. The code generator and selector is shown in Fig. 3. The 4-MHz VCO clocks a binary divider chain called the C-counter, whose stage outputs are labeled $C_{-1}, C_0, C_1, \cdots, C_{24}$. Outputs C_1 to C_{16} ($\hat{c}_1$ to $\hat{c}_{16}$) are the local replicas of the received code sequences, and stages C_{17} to C_{24} are required for system timing. A 4-stage counter, called the K-counter, keeps account of which sequence is being tracked. This counter controls two data selectors whose outputs are $\hat{c}_k$ and $\hat{c}_{k+1}$ when the K-register contains the binary representation of $k + 1$. The reference signals $\hat{w}_k$ and $\hat{w}_{k+1}$ and the transition indication information are simply derived from $\hat{c}_k$, $\hat{c}_{k+1}$, C_1 and C_7.

Whenever the lock detector decides that the received signal has changed from w_k to w_{k+1}, the K-counter is incremented, which has the effect of replacing $\hat{c}_k$ by $\hat{c}_{k+1}$ and $\hat{c}_{k+1}$ by $\hat{c}_{k+2}$. At the same time, if the lock detector decides that the phase of the new received sequence is such that $\hat{c}_{k+1}$ must be inverted, an invert pulse is gated through the demultiplexer to the direct set line of the appropriate flip-flop in the C-counter, under control of the K-counter. Since lock decisions are made when C_{21} switches high, C_{k+1} is zero when this pulse arrives, so that the effect is always to invert $\hat{c}_{k+1}$.

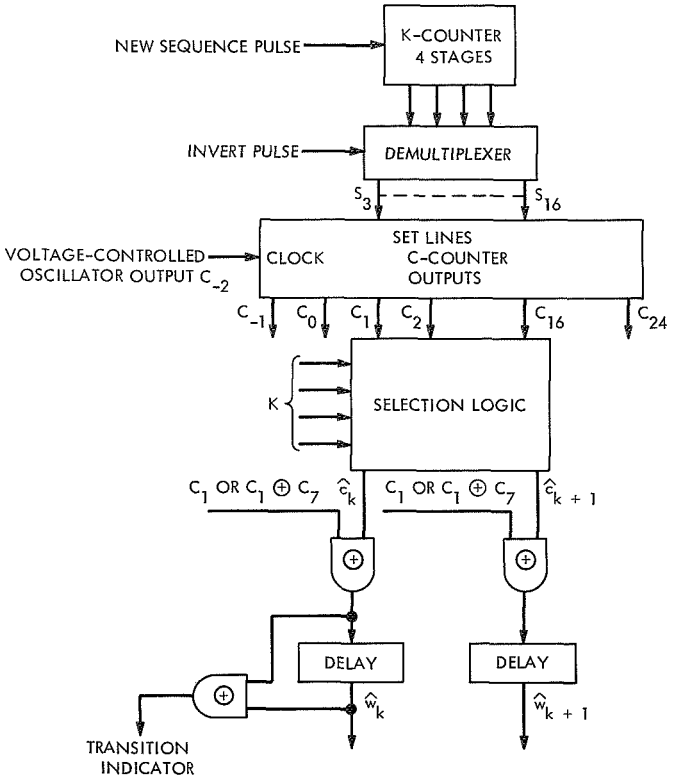

Fig. 3. Code generator and selector

c. Lock detector. The lock detector is shown in Fig. 4. Its functions are: (1) during initial acquisition, to detect when phase lock on w_2 is achieved and to control the loop bandwidth narrowing procedure, (2) during tracking of the various w_k, to determine when the received signal changes to w_{k+1} and the sign of w_{k+1} relative to the local replica, and (3) to determine when the loop is not locked on any component so that the reacquisition can be initiated.

Initial acquisition mode. At the beginning of the acquisition procedure, all registers and counters are set to zero. The two most significant bits of the control (M) counter control whether the clean-up loop is in the acquisition mode or the tracking mode, and the PLL bandwidth. This is shown in Table 1. During acquisition, the A-counter correlates the received signal S with $\hat{w}_2$. Only the samples of S occurring halfway between the transitions of $\hat{c}_1$ are used. The A-counter counts down when S and $\hat{w}_2$ are equal, and up otherwise, and the result is sampled when C_{24} comes high, which is after 2^{23} samples. At this time, A and C_{24} are reset, and the M-counter

Table 1. Control counter states

M_3	M_4	Mode	Bandwidth
0	0	Acquisition	Wide
1	0	Acquisition	Medium
0	1	Acquisition	Narrow
1	1	Tracking	Narrow

is incremented if the final count in A was less than -2^{15}, which is detected by a negative (one) sign bit and at least one *zero* in the most significant eight magnitude bits. When the threshold is exceeded four times, the state of M_3 and M_4 changes, so the bandwidth is narrowed or the mode is switched to tracking. If M_3 does not change within eight trials, it is assumed that the loop is not in lock and the system is reset so that the acquisition procedure starts anew.

Tracking mode. The tracking mode begins when the M-counter reaches $M_3 = M_4 = 1$. The A- and B-counters, respectively, now correlate S with the difference and sum of $\hat{w}_k$ and $\hat{w}_{k+1}$ over intervals of length 2^{20} μsec using the three samples of S taken between each transition of $\hat{c}_1$. In particular, A counts only when $\hat{w}_k \neq \hat{w}_{k+1}$, B counts only when $\hat{w}_k = \hat{w}_{k+1}$, and when counting, both counters count down if $S = \hat{w}_k$. When C_{21} comes high, the signs of the counters are sampled and reset and C_{21} is reset. The counter values at the sampling times are

$$S_A = -\frac{1}{2} \sum_{3 \times 2^{20} \text{ counts}} S(\hat{w}_k - \hat{w}_{k+1}) \tag{2}$$

and

$$S_B = -\frac{1}{2} \sum_{3 \times 2^{20} \text{ counts}} S(\hat{w}_k + \hat{w}_{k+1}) \tag{3}$$

where S, $\hat{w}_k$ and $\hat{w}_{k+1}$ have the values ± 1, and the arithmetic is real arithmetic. The index of summation is not shown. When the summations are completed, both S_A and S_B will most likely be negative if the signal input is w_k. If w_{k+1} is present, S_A will most likely be positive if $w_{k+1} = \hat{w}_{k+1}$ and S_B will be positive if $w_{k+1} = -\hat{w}_{k+1}$. When either is positive, the new sequence pulse is generated and the K-counter is incremented (unless $k = 16$), and C_{k+1} is inverted if $S_B > 0$. At the sampling times, S_A and S_B should never both be positive, although this will occur with probability 1/4 if neither w_k nor w_{k+1} is the input signal. For example, this would occur if the transmitter has returned to w_2. If both S_A and S_B are positive, it is assumed that the loop is out of lock and acquisition is reinitiated.

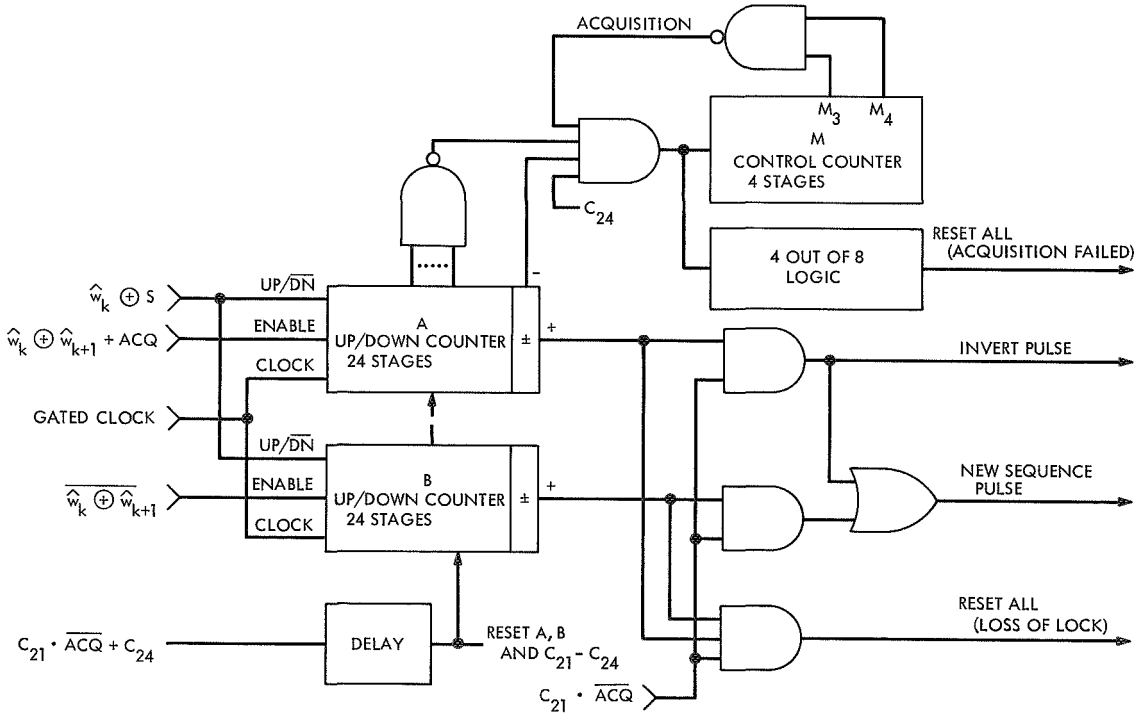

Fig. 4. Lock detector

4. Performance Analysis

The design goal for the clean-up loop is to achieve a high (> 0.99) probability of successful transponding when the ranging signal SNR at the limiter output is -46 dB in a bandwidth of 1.5 MHz. The model assumed for analysis of the system performance was chosen for simplicity of analysis and because this model was easily duplicated for the experimental verification of performance, the results of which are given in *Subsection 5*. The block diagram is shown in Fig. 5. White gaussian noise is added to the signal $w_k(t)$, and the result is passed through an RC low-pass filter with time constant $\alpha = 1/6$ μsec, and then hard limited and sampled at a 4 MHz rate and converted to a logic signal. The low-pass filter has the same 1.5 MHz bandwidth as the *Mariner* transponder.

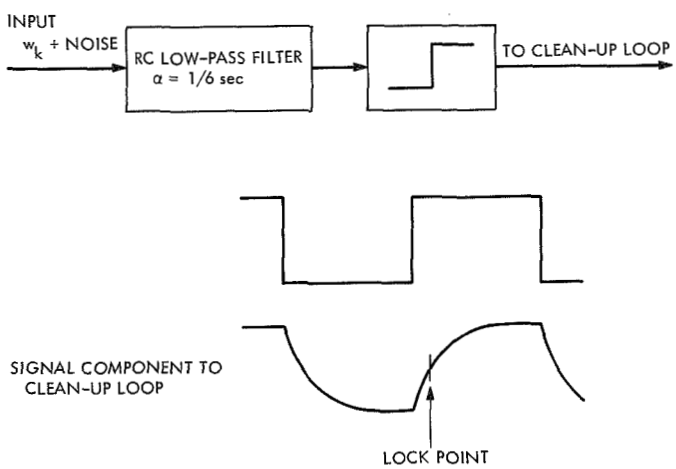

Fig. 5. Model for analysis and experimental tests

At low signal-to-noise ratios, it can be assumed that the signal component of the limiter output has the same shape as at the RC filter output, i.e., piecewise exponential, and that noise is independent of the signal. The noise has unity power and correlation function

$$\rho(\tau) = \frac{2}{\pi} \arcsin\left[\exp\left(-\left|\tau/\alpha\right|\right)\right] \tag{4}$$

The assumed signal waveform is shown with respect to w_k in Fig. 6. Since the RC time constant is short compared to the 1-μsec minimum time between transitions, it can be assumed that the exponential has reached its target voltage, called $\pm V$, by the time a transition occurs. The correct lock point or reference sampling point is assumed to be the point that the waveform crosses zero. Assuming

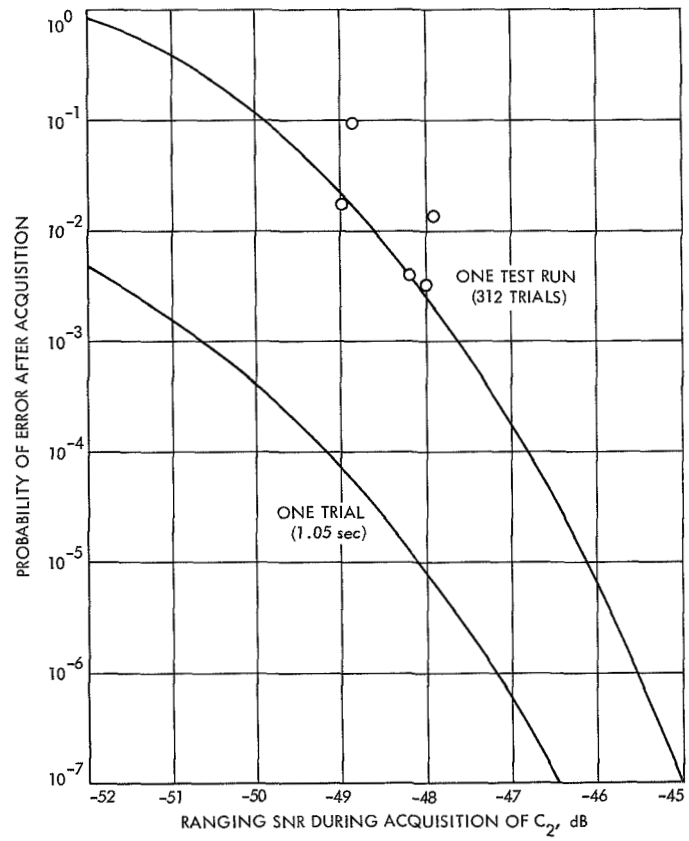

Fig. 6. Probability of error after acquisition

$t = 0$ at the positive zero crossing, the equation of the rising exponential is

$$v(t) = -V + 2V\left[1 - \exp\left(-t/\alpha\right)\right] \tag{5}$$

and the delay from the transition of w_k to the zero crossing is $\alpha \ln(2) = 0.115$ μsec.

The results are presented in terms of R, the SNR at the limiter output when the input is w_2 plus noise, that is, during acquisition. This SNR is

$$R = \frac{5}{6} V^2 \tag{6}$$

For tracking the other components, the SNR is approximately the same as for a 500-kHz square wave input, and is thus approximately

$$R_T = \frac{2}{3} V^2 = \frac{4}{5} R \tag{7}$$

or, in decibels, a 0.96-dB drop from R.

a. Phase detector gain and noise density. The phase detector gain is defined as the derivative with respect to phase of the expected value of the phase detector output. Phase, denoted by ϕ, will always be phase with respect to 500 kHz. The phase detector output is zero when there is no transition, and otherwise has average value equal to the absolute value of $v(t)$ at the phase error in question, i.e., at $t = (\phi/\pi)$ μsec. Letting P_T be the average number of transitions, the phase detector gain is

$$A = \frac{dv(t)}{dt}\frac{dt}{d\phi}\bigg|_{\phi=0}$$

$$= \frac{P_T V}{\pi\alpha} = \frac{1}{\pi\alpha}\left(\frac{6}{5}R\right)^{1/2}$$

$$= 2.09\,R^{1/2}P_T \text{ units/rad} \tag{8}$$

At $R = -46$ dB, $A = 5.24 \times 10^{-3}$ for acquisition and $A = 1.05 \times 10^{-2}$ for tracking.

Because the sampling rate at the phase detector is much less than the Nyquist rate, the noise samples are essentially independent. Hence N_0, the one-sided noise spectral density at zero frequency, is $2P_T \times 10^6$ units^2/Hz.

b. PLL performance. The parameters of the phase-locked loop were chosen to achieve a reasonable loop bandwidth and for convenience in design of the loop filter and DAC connections. For the narrow-loop bandwidth, the parameters are

$$K_1 = 88.0 \text{ rad/sec/unit}$$

and, since the integrator-counter has 23 stages and is updated every microsecond,

$$K_2 = 2^{-23} \cdot 10^6 K_1 = 10.5 \text{ rad/sec}^2\text{/unit}$$

The damping ratio is

$$\zeta = \frac{K_1}{2}(A/K_2)^{1/2} \tag{9}$$

At $R = -46$ dB, $\zeta = 1.0$ for acquisition, and $\zeta = 1.414$ for tracking. Similarly, the one-sided loop bandwidths are given by

$$b_L = \frac{1 + 4\zeta^2}{4}\frac{K_2}{K_1} \tag{10}$$

At $R = -46$ dB, the bandwidth is 0.26 Hz for tracking, and the three bandwidths for acquisition are 2.32, 0.58, and 0.145 Hz.

Steady-state phase error. The steady-state rms error for each case is

$$\sigma_\phi = \left[\frac{N_0 b_L}{A^2}\right]^{1/2} \tag{11}$$

Evaluating this at $R = -46$ dB, σ_ϕ is 0.0725 rad for tracking, and 0.275, 0.138, and 0.0688 rad for acquisition. These correspond to loop SNRs of 22.8 dB for tracking and 11.2, 17.2, and 23.2 dB for acquisition.

Acquisition. The acquisition performance of a phase-locked loop depends on the loop parameters, the phase detector nonlinearity, signal-to-noise ratio, and frequency offset. No adequate theory exists to predict probability of lock as a function of time, or any similar statistic. Heuristically, for a second-order loop one can say that acquisition will occur with high probability in a time less than about $10/b_L$ when the frequency offset does not exceed about $b_L/2$, and when the steady-state loop SNR is high enough so that cycle slipping is infrequent after acquisition. The pertinent parameters during acquisition are very similar for the clean-up loop and for the symbol synchronizer for low SNR coded systems (Ref. 2 and SPS 37-61, Vol. II, pp. 63–68). Both have perfect integrator second-order loops and transition tracking phase detectors (decision directed for the symbol synchronizer), and in both cases b_L is narrowed by a factor of 4 after acquisition. Acquisition time experiments were run for the symbol synchronizer at a symbol SNR of -3 dB, which resulted in a 12-dB loop SNR in the acquisition bandwidth. This compares to a loop SNR of 11.2 dB for the clean-up loop. Two pertinent experimental results for the symbol synchronizer (SPS 37-61, Vol. II) are: (1) that for a frequency offset of $b_L/2$ Hz, acquisition occurs with probability 0.998 within $16/b_L$ sec, and (2) that if the bandwidth is narrowed to one-fourth of its original value $8/b_L$ sec after start of acquisition, the loop will be in lock at that time and remain in lock upon bandwidth narrowing with probability 0.99. Applying these results to the clean-up loop, acquisition should occur with probability 0.99 to 0.998 in 3.5 to 7.0 sec, provided that the frequency offset is less than 1.16 Hz, reflected to the center frequency of 500 kHz.

c. Acquisition lock detection. During initial acquisition, S is sampled at 1-μsec intervals and correlated against $\hat{w}_2$ using sums over 2^{23} samples. Since the sampling rate is well below the Nyquist rate, the variance of the sum is approximately $\sigma^2 = 2^{23}$. When the phase error is zero, samples are taken 0.5 and 1.5 μsec after each transition. Since the time constant is 1/6 μsec, the means of $S\hat{w}_2$ at

these two times are $V(1 - e^{-3})$ and $V(1 - e^{-9})$. The mean of the summation is thus

$$n = 2^{22} V (2 - e^{-3} - e^{-9}) = 1.95 \times 2^{22} V \qquad (12)$$

The summation is compared to a threshold of $T = 2^{15}$, and the probability of not exceeding this threshold when the phase error is zero is

$$P_{\text{MISS}} = 0.5 \operatorname{erfc}\left(\frac{(m - T)}{\sqrt{2}\sigma}\right) \qquad (13)$$

For $R = -46$ dB, $P_{\text{MISS}} \approx 2 \times 10^{-5}$. The probability of failing to exceed the threshold 5 times in eight trials, which would be required for false recycling of the acquisition procedure, is about 10^{-22}.

It is also possible for false alarms to occur, i.e., for T to be exceeded when the loop is not in lock. For this case, the mean of the summation is zero, assuming the loop is slipping cycles. The reason that a correlation time of 2^{23} μsec = 8.4 sec is used is that when the loop is slipping cycles, it will most certainly slip several cycles in this time. If the loop slipped only part of a cycle during the correlation, the mean of the integration could be quite large, depending on the exact phase relationships. However, if the phase error is changing at a rate of less than one cycle in 8.4 sec, the error is most likely approaching zero, i.e., the loop is achieving lock. For these reasons, it is justified to assume that the mean of the correlation is zero when the loop is not in lock. The probability of false alarm during initial acquisition is then

$$P_{FA} = 0.5 \operatorname{erfc}\left(\frac{T}{\sqrt{2}\sigma}\right) = 0.5 \operatorname{erfc}(8) \qquad (14)$$

which is infinitesimal.

d. New sequence detection. After lock is acquired and the loop bandwidth is narrowed, the lock detector makes a decision every 1.05 sec as to whether the received signal w_k is equal to $\hat{w}_k$, $\hat{w}_{k+1}$, or $-\hat{w}_{k+1}$, according to Eqs. (2) and (3). Each decision is called a trial. Because the PLL has no reference signal from the time w_k changes until this change is detected, the rms phase error increases, and it is important to detect changes before a cycle slip can occur. Also, the rms phase error is larger at the instant when the new reference signal is applied than in the steady state. Since this can degrade the lock detector performance, the first 8 trials after detection of a new sequence are inhibited, which places a constraint of about 10 trials or 10.5 sec on the minimum time for which each sequence must be transmitted. After the 8 trial waiting periods, the phase error statistics will have essentially reached steady state.

The probability of a false decision on any trial is the probability that either S_A or S_B has the wrong sign. These events are independent, since half of the terms in S_A are zero and the other half of the terms in S_B are zero. S_A and S_B are also identically distributed, except for the signs of the means. The mean and variance of S_A and S_B were calculated taking into account the phase jitter, assuming the loop to be in the steady state. The mean was determined as a function of phase error, and then averaged over ϕ, assuming ϕ to be gaussian. The variance was calculated by adding the variance of the mean to the variance due to noise, and the probability P_S of having the wrong sign was calculated assuming S_A and S_B to be gaussian. Finally, the probability of an error in a trial is

$$P_E = 1 - (1 - P_S)^2 \approx 2P_S$$

This probability is plotted as a function of R in Fig. 6.

As shown in Fig. 6, the probability of making an error on a given trial is very small, for example, $P_E = 0.18 \times 10^{-7}$ at $R = -46$ dB. This means that the probability is correspondingly high and that a change in w_k will be detected on the first trial which starts after the change occurs. This requires, at most, 2.1 sec. (It can also be shown that the probability of an error on a trial during which a change in w_k occurs is less than P_E.)

It is shown elsewhere (*Section B*) that the increase in mean square phase error after loss of signal at $t = 0$ is

$$\frac{\sigma_\phi^2(t)}{\sigma_\phi^2(t=0)} = 1 + \frac{16\zeta^2}{4\zeta^2 + 1}(2b_L t) + \left(\frac{8\zeta^2 + 1}{4\zeta}\right)\left(\frac{4\zeta}{4\zeta^2 + 1}\right)^3 \times (2b_L t)^2 + \frac{1}{3}\left(\frac{4\zeta}{4\zeta^2 + 1}\right)^4 (2b_L t)^3$$

For the tracking loop parameters, $\sigma_\phi(t = 2.1 \text{ sec}) = 2.4\sigma_\phi(0) = 0.174$ rad with $R = -46$ dB. Since this rms error corresponds to a loop SNR of over 15 dB, it is unlikely that the loop will slip a cycle in this time. We conclude that the new sequence will be detected within 2.1 sec and without a cycle slip, with probability approximately $1 - P_E$.

5. Experimental Verification

An experimental test program has been initiated to verify the theoretical performance of the clean-up loop.

Results to date indicate that the performance after initial acquisition is as anticipated, within the tolerance of the test parameters.

The experimental test setup is shown in Fig. 5. For initial acquisition, component w_2 was generated for about 8 min. Then components w_3 through w_{16} were generated for 32×2^{20} μsec each, or 32 trial intervals. Acquisition was deemed to have occurred if the change from w_2 to w_3 was detected properly. Since detection is inhibited for 8 detection intervals after detection of a new component, there were 312 trials during which an error could occur, 24 on each of 13 components. The theoretical probability of an error in 312 trials is shown in Fig. 6, with the dots denoting experimental results. The dots represent experiments over 428 to 1460 cases, which required approximately a week for each experiment. The results agree with theory to within 1 dB, with the discrepancies attributed to inaccuracy in measuring the RC filter bandwidth and the signal and noise powers, to variations in the noise statistics over the considerable time required for each test point, and to statistical fluctuations.

6. Problem Areas and Suggestions for Improvement

Up to this point we have discussed the clean-up loop as implemented in the prototype model. We now consider some of the problems encountered, and suggest methods for improvement.

The major problem area concerns initial acquisition and the long-term frequency stability of the VCO. Also of major importance is a reduction in the power required by the high-speed digital logic and simplification of the logic. Less important, but still worthy of consideration, is the approximate 2-dB improvement which is attainable by utilizing the received ranging signal before it is hard limited.

a. Frequency stability and acquisition. The uplink ranging frequency is locked to the RF carrier frequency, which is doppler compensated and adjusted to the best lock frequency of the spacecraft RF VCO. As the clean-up loop is implemented, rapid acquisition will occur only when the frequency offset is less than $b_L/2$, or 1.16 Hz out of 500 kHz at the design point SNR. The difference between the center frequencies of the RF and ranging channel VCOs must therefore be maintained to within about ± 2 ppm over the spacecraft temperature range, for the life of the mission, which might be about 10 yr. This is a difficult, if not impossible, requirement to satisfy with independent VCOs. If adequate VCOs cannot be obtained, two possibilities for solving the problem are: (1) to use a more elaborate acquisition procedure, and (2) to reference the ranging channel frequency to the spacecraft RF VCO.

Acquisition procedures. An obvious way to enable acquisition over a wider range of frequency offsets is to sweep the VCO. This could be achieved quite simply by inserting counts into the integrator up/down counter in the linear filter, thereby sweeping the VCO control voltage. After acquisition is detected, the sweeping would be disabled. The major disadvantage to this procedure is that the first ranging component would have to be transmitted for a long enough time for the frequency to be swept across the band of possible offsets.

A second procedure might be preferable if offsets are known to occur mainly because of aging, rather than because of shorter term effects, like temperature drift. If this is the case, the center frequency could be set at the start of acquisition to the last known lock frequency. This frequency is determined by the value in the integrator-counter the last time the clean-up loop was known to be in lock. The disadvantage to this procedure is that the VCO might be set to the wrong initial frequency if the loop was really out of lock the last time it was thought to be in lock. This disadvantage may not be too serious, however, because recovery could be accomplished by transmitting the first component for a long enough time for acquisition to occur even with a high-frequency offset.

Referencing to RF VCO. The major disadvantage to using the RF VCO frequency as a reference for the ranging channel is that it makes the operation of the clean-up loop dependent on the configuration and operation of the rest of the spacecraft. This is not desirable, for one would like to have a mission independent ranging transponder. A second disadvantage is that cycle slipping in the RF phase-locked loop would cause phase changes in the ranging PLL. This is a less serious disadvantage, because ranging is otherwise dependent on accurate carrier lock.

Two ways to utilize a reference from the RF VCO are shown in Fig. 7. In both cases, f_r is the required ranging clock frequency, and $f_r \pm f_c$ is a frequency derived from the RF VCO, and therefore precisely related to the ranging frequency. The required ranging frequency f_r is obtained by mixing $f_r \pm f_c$ with f_c. The first method (Fig. 7a) is a standard mixer-modulator VCO, with the higher frequency having a precise reference. Since the VCO frequency, f_c, is fairly low, it can have a center frequency tolerance which is fairly high percentage of its center frequency, and still be a small percentage of f_r.

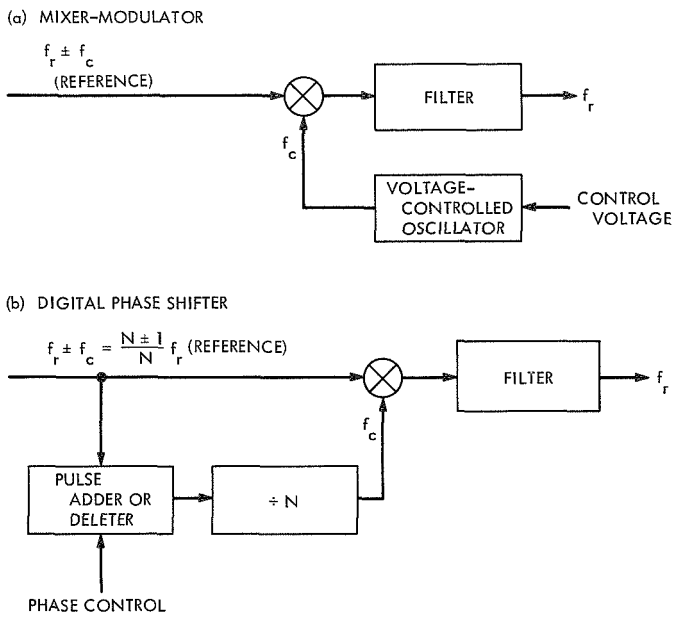

Fig. 7. Use of RF VCO reference

The remainder of the PLL could be unchanged. The second method, Fig. 7b, uses a digital phase shifter rather than a VCO. In this case, f_c is f_r/N for some convenient integer N. Since f_c is derived from f_r, its frequency is exact. Phase control of f_c is accomplished by digitally phase shifting f_c by adding or deleting pulses into the frequency divider. The phase is changed by $2\pi/N$ each time a count is added or deleted. This method has the advantage that the frequencies are exact, so that a first-order PLL is sufficient. Furthermore, no VCO or DAC is required, since the system is entirely digital.

b. Power reduction. The prototype implementation of the clean-up loop was not designed to minimize power. All of the logic circuits are high-speed transistor-transistor logic even where the speed is not required, and the 80 odd integrated circuits consume about 12 W. It is obvious that some power can be saved by using slower circuits where feasible, but greater savings are achievable by reducing the need for high-speed circuits.

The circuits which consume the most power are the up/down counters. In the case of the lock detectors, this problem can be overcome by using up counters only, after presetting to the proper value. This is possible because the total number of up and down counts during each detection interval is known, so that sufficient information is contained in the up counts only. In the case of the integrator-counter in the linear filter, the speed requirements could be reduced by updating the counter less often, say every 16 μsec rather than every 1 μsec. This would also eliminate one 4-bit counter chip, and would enable use of an integrated circuit DAC rather than the larger DAC now utilized because of its higher switching speed. The reduction in sampling rate can be accomplished by proper modifications to the nonlinear filter.

c. Use of non-limited input signal. The input signal to the clean-up loop is now assumed to be hard limited, but about a 2-dB gain can be realized by using a non-limited signal. This would enable the lock detection interval to be reduced and the PLL bandwidth to be increased, at the expense of some complication in the mechanization. First, an automatic gain control (AGC) system and an analog-to-digital converter would be required. For the PLL, it would be necessary to modify the filters to accept multilevel inputs. One way to do this would be to modify the nonlinear filter so as to have multilevel inputs and binary outputs. Modifications to the lock detectors would be more complicated. Adders with up/down counters for overflow would be required rather than up counters only, and selection of a threshold for initial lock detection would be difficult because of inaccuracies in any AGC system. Overall, it appears that the possible 2-dB improvement in performance would not justify the increased complexity of the clean-up loop.

References

1. Martin, W. L., "A Binary-Coded Sequential Acquisition Ranging System," in *Proceedings of the Third International Conferences in System Sciences*, University of Hawaii, Honolulu, January 1970.

2. Hurd, W. J., and Anderson, T. O., "Digital Transition Tracking Symbol Synchronizer for Low SNR Coded Systems," *IEEE Trans. Commun. Technology*, Vol. COM-18, April 1970, pp. 141–147.

B. Performance of a Phase-Locked Loop During Loss of Signal, *W. J. Hurd*

1. Introduction

The signal input to a phase-locked loop (PLL) is sometimes lost after the loop has reached steady-state conditions, as can occur in the article on ranging clean-up loops (*Section A*). This loss may be accidental, due to fading, for example, or it may be deliberate due to a change in the transmitted signal which causes the received signal to be orthogonal to all phase shifts of the loop reference signal. In either case, it may be desirable for the loop to remain in lock with the original input signal.

Timor (SPS 37-61, Vol. III, pp. 33–37) has considered the first passage problem for the phase error of a first-order PLL when the signal is lost, the major result being the probability that a cycle slip occurs up to time t after loss of signal. In most applications, however, the probability distribution of phase error is of more interest, for the closest lock point is the correct one whenever the absolute phase error is less than π, regardless of whether or not complementing cycle slips have occurred. Here we determine the joint probability distribution of phase error and all other state variables for any order PLL, conditioned on the state of the loop at the time of loss of signal. When the linearized model for the loop is valid, which is usually the case when few cycle slips are allowed, we determine the unconditional distribution of the state variables.

The increase in phase error variance as a function of time after loss of signal is shown graphically for the first-order loop with and without frequency offset and for the perfect integrator second-order loop with various damping ratios.

2. Model

The PLL model used is shown in Fig. 1. The noise $n(t)$ is assumed to be white with one-sided spectral density N_0; $\phi(t)$ is the phase error, and Ω_0 is the offset between the frequency of the input signal and the voltage-controlled oscillator (VCO) center frequency. The sign of Ω_0 is chosen so that positive Ω_0 causes a positive change in ϕ. The phase detector nonlinearity $g(\phi)$ has slope $A = g'(0)$ at $\phi = 0$, so the open-loop transfer function of the linearized model for the loop is $AF(s)/s$. Note that $g(\phi)$ is dependent on the amplitude of the input signal, and is zero for all ϕ when the input signal is zero or is orthogonal to all phase shifts of the VCO output signal.

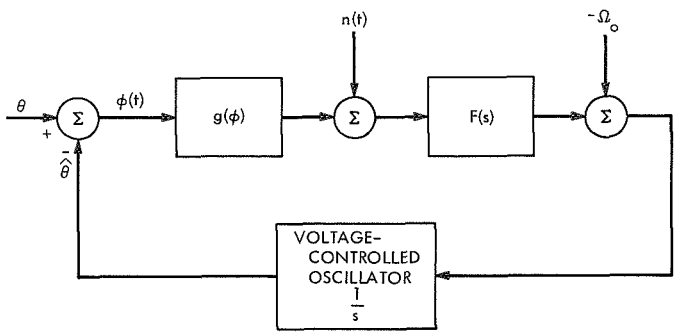

Fig. 1. Model for phase-locked loop

3. Problem and General Solution

Suppose that $F(s)$ is the transfer function of an $(N-1)$th order linear system, so that the loop is Nth order. The state of the loop is completely described by $\phi(t)$ and $N-1$ other state variables. These state variables should be judiciously chosen so as to have no white noise components so that the variances are finite. Suppose that the signal input to the loop is lost at time $t = 0$, i.e., $g(\phi) = 0$ for $t > 0$. We obtain the means and cross-correlation functions of all state variables for $t \geqq 0$, conditioned on the initial conditions of the state variables at $t = 0$. If $n(t)$ is gaussian, we thus obtain the conditional distribution of the state variables. Furthermore, if the loop has reached steady state by time $t = 0$ and the loop signal-to-noise ratio is high (so that the linear approximation $g(\phi) = A\phi$ is valid) then the state variables are jointly gaussian at $t = 0$ and we can obtain the unconditional distribution of the state variables for $t > 0$.

Observe that setting A equal to zero at $t = 0$ is equivalent to opening the loop at the output of the phase detector nonlinearity $g(\phi)$. The response of the system is then just the response of the open-loop system to $n(t)$ plus the initial conditions. Let

$$X = [x_1(t), \cdots, x_N(t)]$$

be the vector of the state variables, with $\phi(t) = x_1(t)$. Because the system is linear, we can write

$$X = Y + W \tag{1}$$

where $Y = (y_1, \cdots, y_N)$ is the portion of X due to the initial conditions and frequency offset and $W = (w_1, \cdots, w_N)$ is the portion due to the noise $n(t)$.

Since $n(t)$ is zero mean, so are the $w_i(t)$. Hence, conditioned on $Y(t = 0)$ the mean of $x_i(t)$ is $y_i(t)$, and the covariance function of x_i and x_j is the covariance of w_i

and w_j. Given $Y(t=0)$, $Y(t)$ can be evaluated for $t \geqq 0$ by several well-known techniques.

To evaluate the covariances, let $h_i(t)$ be the weighting function relating to $w_i(t)$ to $n(t)$. In other words, if $H_i(s)$ is the transfer function from $n(t)$ to $w_i(t)$, then $h_i(t)$ is the inverse Laplace transform of $H_i(s)$. We can then write

$$w_i(t) = \int_0^t n(\tau)\, h_i(t-\tau)\, d\tau \tag{2}$$

The covariance of w_i and w_j is simply

$$\begin{aligned} R_{ij}(t_i, t_j) &= E\{w_i(t_i)\, w_j(t_j)\} \\ &= \int_0^{t_i}\int_0^{t_j} h_i(t_i-\alpha)\, h_j(t_j-\beta) \\ &\quad \times E\{n(\alpha)\, n(\beta)\}\, d\beta\, d\alpha \end{aligned} \tag{3}$$

Since

$$E\{n(\alpha)\, n(\beta)\} = \frac{N_0}{2}\delta(\alpha-\beta)$$

the integral with respect to β is equal to

$$\frac{N_0}{2} h_j(t_j-\alpha)$$

for $0 \leqq \alpha \leqq t_j$ and zero elsewhere, so

$$R_{ij}(t_i, t_j) = \frac{N_0}{2}\int_0^{\min(t_i, t_j)} h_i(t_i-\alpha)\, h_j(t_j-\alpha)\, d\alpha \tag{4}$$

For $t = t_i = t_j$, this specializes to

$$\begin{aligned} R_{ij}(t,t) &= \frac{N_0}{2}\int_0^t h_i(t-\alpha)\, h_j(t-\alpha)\, d\alpha \\ &= \frac{N_0}{2}\int_0^t h_i(\alpha)\, h_j(\alpha)\, d\alpha \end{aligned} \tag{5}$$

4. First-Order Loop

For the first-order PLL, let $F(s) = K$ and denote the steady-state probability density of phase error by $p_s(\phi)$. If the linearized model is valid, $p_s(\phi)$ is gaussian with mean equal to the static phase error ω_0/AK, and variance

$$\sigma_s^2 = \frac{N_0 K}{4A} = \frac{N_0 b_L}{A^2} \tag{6}$$

where b_L is the one-sided loop bandwidth, $AK/4$. The density $p_s(\phi)$ is also well known when the phase detector nonlinearity is sinusoidal, i.e., $g(\phi) = A\sin\phi$, and is given by Viterbi (Ref. 1) and elsewhere.

To determine the statistics of $\phi(t)$ after the signal is lost at $t=0$, let $\phi_0 = \phi(0)$ and let $\phi_n(t) = w_1(t)$ be the component of $\phi(t)$ due to $n(t)$ for $t > 0$. Then

$$\phi(t) = \phi_0 + \Omega_0 t + \phi_n(t) \tag{7}$$

and

$$\phi_n(t) = -K\int_0^t n(t)\, dt \tag{8}$$

The noise term $\phi_n(t)$ is independent of ϕ_0, since the noise is white, and has zero mean and variance $\sigma_n^2 = K^2 N_0 t/2$. It is, of course, gaussian if $n(t)$ is, so the conditional distribution of $\phi(t)$ is

$$p(\phi(t)\,|\,\phi_0) = \frac{1}{(2\pi)^{1/2}\sigma_n}\exp\left[\frac{-(\phi-\phi_0-\Omega_0 t)^2}{2\sigma_n^2}\right] \tag{9}$$

The unconditional distribution of $\phi(t)$ for $t > 0$ is

$$p_t(\phi) = \int_{-\infty}^{\infty} p(\phi(t)\,|\,\phi_0)\, p_s(\phi_0)\, d\phi_0 \tag{10}$$

and, when $p_s(\phi_0)$ is gaussian, $\phi(t)$ is gaussian with mean

$$\frac{\Omega_0}{AK} + \Omega_0 t \tag{11}$$

and variance

$$\sigma_t^2 = \sigma_s^2 + \sigma_n^2 = \frac{N_0 K}{4A}(1 + 2AKt) \tag{12}$$

It is interesting to rewrite this expression in terms of the two-sided loop bandwidth w_L,

$$\sigma_t^2 = \frac{N_0 w_L}{2A^2}(1 + 4w_L t) \tag{13}$$

When the signal is removed, the variance of phase error doubles in one fourth of an inverse loop bandwidth, or one half of one loop time constant.

5. Second-Order Loop

For a second-order loop we can find the unconditional probability density of ϕ (and of the other state variable) only when the loop is linearized. This is because the joint

density of the two state variables in the steady state is not known, even though $p_s(\phi)$ is known for some $g(\phi)$.

Suppose the loop filter is

$$F(s) = K\frac{1+\tau_2 s}{1+\tau_1 s} \tag{14}$$

If we let $K_2 = K/\tau_1$ and $K_1 = K_2\tau_2$, then $F(s)$ reduces to the filter for a perfect integrator second-order loop as $\tau_1 \to \infty$ with K_1 and K_2 constant,

$$F(s) = K_1 + K_2/s, \qquad \tau_1 = \infty \tag{15}$$

Appropriate state variables can be chosen by factoring $F(s)/s$ into

$$\frac{F(s)}{s} = \frac{K_2}{s+1/\tau_1}\left[\frac{K_1}{K_2} + \frac{1}{s}\right] \tag{16}$$

and considering the factors of $F(s)$ to be implemented in series, with the $K_2/(s+1/\tau_1)$ factor first, as shown in Fig. 2. We then chose ϕ and the negative of the output of the first factor, ω, as the state variables. This choice is appropriate both because ω has no white noise component and because $E\{\omega\} = -\Omega_0$ for the perfect integration case.

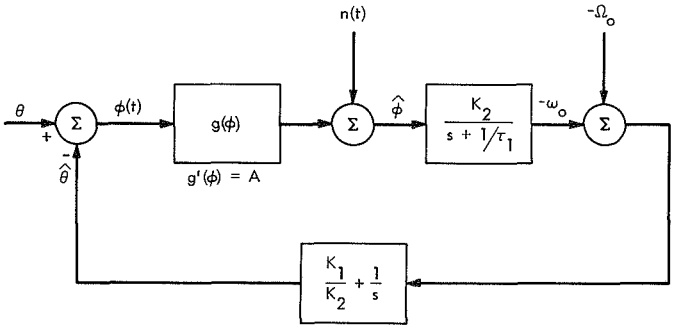

Fig. 2. Model for second-order loop

In the formulation of Eq. (1),

$$X(t) = (\phi(t), \omega(t)) = (x_1(t), x_2(t)) = Y + W$$

where Y is the portion of X due to initial conditions and Ω_0, and W is the portion due to noise for $t > 0$.

a. Conditional distribution. The components of ϕ and ω due to Ω_0 and the initial conditions can be obtained by inspection. For the imperfect integrator, $\tau_1 < \infty$,

$$y_2[t|Y(0)] = \omega_0 \exp(-t/\tau_1) \tag{17}$$

and

$$y_1[t|Y(0)] = \frac{K_1}{K_2}[\Omega_0 + y_2(t)] + \int_0^t [\Omega_0 + y_2(\tau)]\,d\tau + C$$

$$= \phi_0 + \Omega_0 t + \omega_0\left(\tau_1 - \frac{K_1}{K_2}\right)[1 - \exp(-t/\tau_1)] \tag{18}$$

For the perfect integrator, $\tau_1 = \infty$,

$$y_2[t|Y(0)] = \omega_0 \tag{19}$$

and

$$y_1[t|Y(0)] = \phi_0 + (\Omega_0 + \omega_0)t \tag{20}$$

Note that y_1 and y_2 are the means of ϕ and ω for $t > 0$, conditioned on ϕ_0 and ω_0.

To determine the conditional variances of ϕ and ω, we evaluate the variances of $w_1(t)$ and $w_2(t)$, the components of $\phi(t)$ and $\omega(t)$, respectively, due to white noise after $t = 0$. For the imperfect integrator case, the weighting functions from $n(t)$ to w_2 and w_1 are

$$h_2(\tau) = u(\tau) K_2 \exp(-\tau/\tau_1) \tag{21}$$

where $u(\tau)$ is the unit step function, and

$$h_1(\tau) = u(\tau) K_2\left[\frac{1}{\tau_1} + \left(1 - \frac{1}{\tau_1}\right)\exp(-\tau/\tau_1)\right] \tag{22}$$

Similarly, for $\tau_1 = \infty$,

$$h_2(\tau) = K_2 u(\tau) \tag{23}$$

and

$$h_1(\tau) = u(\tau)[K_1 + K_2 t] \tag{24}$$

Applying Eq. (5), the variance of ω conditioned on ϕ_0 and ω_0 is

$$\sigma_2^2(t) = \frac{N_0 K_2^2 \tau_2}{4}[1 - \exp(-2t/\tau)], \qquad \tau_1 < \infty \tag{25}$$

$$= \frac{N_0 K_2^2 t}{2}, \qquad \tau_1 = \infty \tag{26}$$

and the variance of ϕ conditioned on ϕ_0 and ω_0 is

$$\sigma_1^2(t) = \frac{N_0K_2^2}{2}\left(\frac{t}{\tau_1^2} + \left(1 - \frac{1}{\tau_1}\right)[1 - \exp(-t/\tau_1)] + \frac{(\tau_1 - 1)^2}{2\tau_1}[1 - \exp(-2t/\tau_1)]\right), \quad \tau_1 < \infty \tag{27}$$

$$= \frac{N_0}{2}\left[K_1^2 t + K_1K_2t^2 + \frac{K_2^2}{3}t^3\right], \quad \tau_1 = \infty \tag{28}$$

Similarly, the conditional covariance of $\phi(t)$ and $\omega(t)$ is

$$\sigma_{12}^2(t) = \frac{N_0K_2^2}{2}\left[\exp(-t/\tau_1) + \left(\frac{\tau_1 - 1}{2}\right)\exp(-2t/\tau_1)\right], \quad \tau_1 < \infty \tag{29}$$

$$= \frac{N_0}{2}(K_1K_2t + \tfrac{1}{2}K_2^2t^2), \quad \tau_1 = \infty \tag{30}$$

At time t, the state variables ϕ and ω, conditioned on ϕ_0 and ω_0, are jointly gaussian with means y_1 and y_2, variances σ_1^2 and σ_2^2, and covariance σ_{12}^2.

b. Unconditional distributions. To obtain the unconditional distributions of $\phi(t)$ and $\omega(t)$, we must utilize the distributions of ϕ_0 and ω_0. The means are given by Eqs. (17) to (20), so it remains only to obtain the second moments when the linear PLL model is assumed and ϕ_0 and ω_0 are jointly gaussian.

In terms of the damping ratio ζ, and the loop natural frequency β, the closed-loop transfer functions from $n(t)/A$ to ϕ and ω, respectively

$$F_1(s) = \frac{\left(2\zeta\beta - \frac{1}{\tau_1}\right)s + \beta^2}{s^2 + 2\zeta\beta s + \beta^2} \tag{31}$$

and

$$F_2(s) = \frac{\beta^2 s}{s^2 + 2\zeta\beta s + \beta^2} \tag{32}$$

where

$$2\zeta\beta = AK_1 + 1/\tau_1$$

and

$$\beta^2 = AK_2$$

Using the steady state linear theory, the second moments are obtained by integrating the (cross) spectral densities, e.g.,

$$E\{\phi_0\omega_0\} = \frac{N_0}{2A^2}\int_{-\infty}^{\infty} F_1(j\omega)\,F_2^*(j\omega)\frac{d\omega}{2\pi} \tag{33}$$

We present further details only for the perfect integrator case, $\tau_1 = \infty$. For this case,

$$E\{\phi_0^2\} = \frac{N_0}{2A^2}\frac{\beta}{2}\left(2\zeta + \frac{1}{2\zeta}\right) \tag{34}$$

$$E\{\omega_0^2\} = \frac{N_0}{2A^2}\frac{\beta^3}{4\zeta} \tag{35}$$

and

$$E\{\phi_0\omega_0\} = \frac{N_0}{2A^2}\frac{\beta^2}{2} \tag{36}$$

The two-sided loop bandwidth is

$$w_L = \frac{\beta}{2}\left(2\zeta + \frac{1}{2\zeta}\right)\text{Hz} \tag{37}$$

The moments of Y can now be obtained using Eqs. (19) and (20). The means are zero, since $E\{\omega_0\} = -\Omega_0$. The second moments are

$$E\{y_1^2\} = E\{(\phi_0 + \omega_0 t)^2\} = \frac{N_0}{2A^2}\left[w_L + \beta^2 t + \frac{\beta^3}{4\zeta}t^2\right] \tag{38}$$

$$E\{y_2^2\} = E\{\omega_0^2\} \tag{39}$$

and

$$E\{y_1y_2\} = E\{\phi_0\omega_0\} + E\{\omega_0^2\}\,t \tag{40}$$

To obtain the moments of

$$X = [\phi(t), \omega(t)]$$

we need merely add the moments of W and Y. In particular, the variance of $\phi(t)$ is

$$\sigma_\phi^2(t) = \frac{N_0}{2A^2}\left[w_L + (4\zeta^2 + 1)\beta^2 t + \left(2\zeta + \frac{1}{4\zeta}\right)\beta^3t^2 + \frac{\beta^4}{3}t^3\right] \tag{41}$$

Since $\beta = 4\zeta w_L/(4\zeta^2 + 1)$, we can express the increase in the variance of ϕ with time as

$$\frac{\sigma_\phi^2(t)}{\sigma_\phi^2(0)} = 1 + \frac{16\zeta^2}{4\zeta^2+1} w_L t + \frac{8\zeta^2+1}{4\zeta}\left(\frac{4\zeta}{4\zeta^2+1}\right)^3 (w_L t)^2 + \frac{1}{3}\left(\frac{4\zeta}{4\zeta^2+1}\right)^4 (w_L t)^3 \quad (42)$$

6. Comparison of Phase Error for First- and Second-Order Loops

The variance of phase error at time t sec after loss of signal is given by Eq. (12) for the first-order loop and by Eq. (42) for the perfect integrator second-order loop. To obtain the mean square phase error, we must add the square of the static phase error to the variance. This static phase error is zero for the second-order case, and is Ω_0/AK for the first-order case. To present comparative results, it is convenient to normalize the rms phase error by $\sigma_\phi(0)$, which eliminates dependence on N_0, A, and w_L. Since there is no static error for the second-order case, the rms phase error, normalized by $\sigma_\phi(0)$, depends only on t and ζ. For the first-order case, it depends only on t and Ω_0.

Figure 3 shows the rms phase error, normalized by $\sigma_\phi(0)$, for the second-order loop with several damping ratios and for the first-order loop with no frequency offset and with offset

$$\Omega_0 = (N_0 w_L)^{1/2} \frac{w_L}{2A}$$

This particular frequency offset was chosen because it corresponds to the case when w_L is optimized for given N_0, A, and Ω_0, in the sense that rms phase error is minimized when the signal is present. The rms phase error at

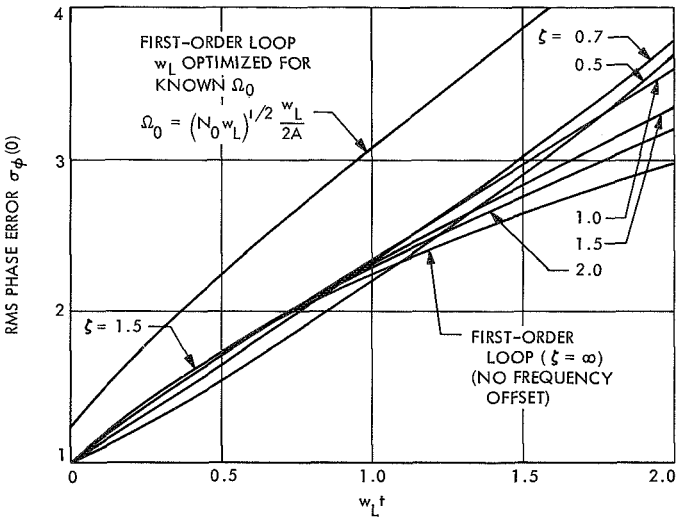

Fig. 3. RMS phase error versus time after loss of signal for second-order phase-locked loop

$t = 0$ is thus as small as is possible with a first-order loop. When there is no frequency offset, there is little difference in performance between the first- and second-order loops for small $w_L t$, but the first-order loop appears to be better for large $w_L t$. This is deceptive, however, for when there is any frequency offset, the optimum bandwidth first-order loop is worse than a second-order loop of the same bandwidth. It is interesting that the damping ratio of the second-order loop is not an important parameter, so that its selection would normally be determined by other considerations in the design of a system.

Reference

1. Viterbi, A. J., "Principles of Coherent Communication," Chap. 4. McGraw-Hill Book Co., New York, 1966.

C. Optimum Configurations for PSK/PM Systems,

U. Timor

1. Introduction

Let $\hat{a}_1(t)_1 \cdots, \hat{a}_N(t)$ be N binary data sequences to be transmitted over a single coherent RF channel by phase modulating the carrier with biphase-modulated squarewave subcarriers (phase-shift keyed/phase modulated). The transmitted signal is

$$s(t) = \sqrt{2P}\sin[\omega_c t + \theta(t)] \tag{1}$$

where P is the total power and $\theta(t)$ is a 2^N-level function of the data according to the type of modulation scheme used. Two methods previously investigated are:

a. Conventional method (Ref. 1). This method uses a linear combination of the data streams, i.e.,

$$\theta(t) = \sum_{k=1}^{N} \theta_k \hat{a}_k(t)\, sq(\omega_k t) = \sum_{k=1}^{N} \theta_k a_k(t) \tag{2}$$

where θ_k are the modulation angles, $sq(\omega_k t)$ are squarewave subcarriers of frequency ω_k, and $a_k(t) = \hat{a}_k(t)\, sq(\omega_k t)$.

b. Interplex method (SPS 37-62, Vol. III, pp. 57–60; and SPS 37-63, Vol. III, pp. 34–37). This method performs some intermultiplications of data sequences to yield the modulating angle

$$\theta(t) = \theta_1 a_1(t) + a_1(t) \sum_{k=2}^{N} \theta_k a_k(t) \tag{3}$$

Comparing these two modulation schemes, it was shown (SPS 37-65, Vol. III, pp. 27–31) that the Interplex method out-performs the conventional one when N is small, while for large N, the power allocation to the data channels determines which configuration is superior.

The above methods are but two of many possible configurations of generating the 2^N-level phase $\theta(t)$. The question arises whether, under some given circumstances, there is another scheme which out-performs both the conventional and Interplex methods. We therefore consider the general modulation angle

$$\theta(t) = \sum_{k=1}^{N} \theta_k x_k(t) \tag{4}$$

where $x_k(t)$ are binary waveforms generated by the data sequences $a_k(t)$. Clearly the conventional scheme $x_k(t) = a_k(t)$, and the Interplex method $x_1(t) = a_1(t)$, $x_k(t) = a_1(t)\, a_k(t)\, k \geqq 2$, are special cases of Eq. (4).

It will be shown that for all N and all possible power allocation to the data and RF channels, either the Interplex or the conventional scheme are optimum. This optimum might not be unique; that is, some other configuration could have the same efficiency. However, no other configuration could out-perform both the conventional and Interplex system. Therefore, to find the optimum performance, it is enough to consider those two schemes.

2. System Optimization

Let the transmitted signal be

$$S(t) = \sqrt{2P}\sin\left(\omega t + \sum_{k=1}^{N} \theta_k x_k(t)\right) \tag{5}$$

where $x_k(t)$ are binary waveforms generated by the data.

Correlating the received signal with a coherent reference ($\sqrt{2}\cos\omega t$) and its quadrative ($\sqrt{2}\sin\omega t$), and performing subcarrier demodulation and filtering, we could separate any of the $2^N - 1$ signals of the form

$$y_{\underset{\sim}{\lambda}}(t) = \prod_{k=1}^{N} [x_k(t)]^{\lambda_k} \tag{6}$$

where

$$\underset{\sim}{\lambda} = \{\lambda_1, \cdots, \lambda_N\}$$

is a binary vector ($\lambda_k = 0$ or 1) and can assume all $2^N - 1$ non-zero combinations. The relative power of the signal $y_{\underset{\sim}{\lambda}}(t)$ is given by

$$P_{\underset{\sim}{\lambda}}(t) = \prod_{k=1}^{N} (\sin^2\theta_k)^{\lambda_k} (\cos^2\theta_k)^{(1-\lambda_k)} \tag{7}$$

Let $\alpha_k = P_k/P_1$, $k = 1, \cdots, N$ be the required ratio of powers in the kth channel to the first one, and let $\tilde{P}_{RF}$ be the minimum RF power required to track the frequency and phase.

An allowable configuration C of the system is defined as any collection of N independent signals from the $2^N - 1$ possible signals $y_{\underset{\sim}{\lambda}}(t)$ and corresponding modulation angles $\theta_1, \cdots, \theta_N$ which satisfy the required ratios α_k, $k = 1, \cdots, N$ and which yield an RF power greater than or equal to $\tilde{P}_{RF}$.

An optimum configuration $\hat{C} = \{\hat{y}_k(t), \hat{\theta}_k, k = 1, \cdots, N\}$, not necessarily unique, is one which maximizes the total data power among all allowable configurations.

Without loss of generality we can assume that:

$$\left.\begin{aligned} &\theta_k \leq \frac{\pi}{2}, \qquad k = 1, \cdots, N \\ &\theta_1 = \max_k \theta_k \end{aligned}\right\} \tag{8}$$

The optimum configurations are given by the following theorem.

THEOREM 1. *Let*

$$\hat{C} = \{\hat{y}_k(t), \hat{\theta}_k, k = 1, \cdots, N\}$$

be an optimum configuration satisfying the required RF power and data power ratios.

Then $\hat{C}$ is either identical or equivalent (with respect to power allocation) to

$$\tilde{C} = \{\tilde{y}_k(t), \tilde{\theta}_k, k = 1, \cdots, N\}$$

where $\tilde{C}$ is given by

$$\begin{aligned} &\tilde{y}_1(t) = x_1(t) \\ &\tilde{y}_k(t) = x_k(t) \qquad \text{if } \hat{\theta}_1 \leq \frac{\pi}{4} \\ &\quad = x_1(t)\, x_k(t) \qquad \text{if } \hat{\theta}_1 > \frac{\pi}{4} \\ &\qquad k = 2, \cdots, N \end{aligned} \tag{9}$$

The proof is given in *Paragraph 3.*

To obtain estimates of the data sequences at the receiver we assign

$$\tilde{y}_k(t) = a_k(t), \qquad k = 1, \cdots, N \tag{10}$$

From Theorem 1, we obtain the modulating binary signals:

$$\begin{aligned} &x_1(t) = \tilde{y}_1(t) = a_1(t) \\ &x_k(t) = a_k(t), \qquad \hat{\theta}_1 \leq \frac{\pi}{4} \\ &\quad = a_1(t)\, a_k(t), \qquad \hat{\theta}_1 > \frac{\pi}{4} \end{aligned} \tag{11}$$

We thus have the following result:

THEOREM 2. *Given the data power ratios α_k $(k=1, \cdots, N)$ and the required RF power, either the conventional system or the Interplex system is optimal. Other configurations might achieve the same efficiency but could never out-perform both of them.*

3. Proof of Theorem 1

a. Case I: $\hat{\theta}_1 \leq \pi/4$

From Eq. (8) we have

$$\hat{\theta}_k \leq \frac{\pi}{4} \qquad k = 1, \cdots, N, \tag{12}$$

and therefore

$$\cos^2 \hat{\theta}_k \geq \sin^2 \hat{\theta}_k \qquad k = 1, \cdots, N \tag{13}$$

Each term $\sin^2 \theta_k$ in $P_{\underset{\sim}{\lambda}}$ (Eq. 7) indicates the presence of $x_k(t)$ in $y_{\underset{\sim}{\lambda}}(t)$ (Eq. 6), and therefore must appear at least once in $P_1, \cdots, P_N$ (otherwise the N signals

$$\hat{y}_1(t), \cdots, \hat{y}_N(t)$$

will be dependent). However, if for some k, $\sin^2 \theta_k$ appears more than once, we can substitute it with $\cos^2 \theta_k$ and thus increase the total data power (Eq. 13) while preserving signal independence. It can be shown that due to the continuity of the θ_k's, the increase in data power can be redistributed to preserve the required power ratios. Thus the data power is maximized if each $\sin^2 \theta_k$ appears once and only once to yield:

$$\left.\begin{aligned} &\tilde{y}_k(t) = x_k(t) \qquad k = 1, \cdots, N \\ &P_k = \sin^2 \theta_k \prod_{i=1}^{N} \cos^2 \theta_i \end{aligned}\right\} \tag{14}$$

Moreover if $\theta_1 < \pi/4$ the optimum is unique

QED. Case I

b. Case II: $\hat{\theta}_1 > \pi/4$

LEMMA 1

If $\hat{\theta}_1 > \pi/4$ than there exists an optimum configuration satisfying:

$$\hat{y}_k(t) = x_1(t)\, z_k(t), \qquad k = 1, \cdots, N \tag{15}$$

where $z_k(t)$ is independent of $x_1(t)$.

Proof

Suppose this is not the case. Then for every optimum configuration there exists some $\tilde{y}_j(t)$ independent of $x_1(t)$. Hence

$$P_j = \cos^2\theta_1 A_j(\theta_2, \cdots, \theta_N) \tag{16}$$

since $\sin^2\theta_1 > \cos^2\theta_1$, the signal

$$u(t) = x_1(t)\,\tilde{y}_j(t) \tag{17}$$

has a power

$$P_u = \sin^2\theta_1 A_j(\theta_2, \cdots, \theta_N) > P_j \tag{18}$$

Therefore $u(t)$ belongs to the optimum set, i.e., $u(t) = \tilde{y}_l(t)$, otherwise the data power would be increased by substituting $\tilde{y}_j(t)$ with $u(t)$.

From Eq. (16) it is evident that $A_j(\theta_2, \cdots, \theta_N)$ must include some $\sin^2\theta_i$ term which we can assume without loss of generality to be $\sin^2\theta_2$. Thus

$$\begin{aligned} P_j &= \cos^2\theta_1 \cdot \sin^2\theta_2 B(\theta_3, \cdots, \theta_N) \\ P_l &= \sin^2\theta_1 \cdot \sin^2\theta_2 B(\theta_3, \cdots, \theta_N) \end{aligned} \tag{19}$$

and

$$P_l > P_j \tag{20}$$

Consider the signal $v(t)$ associated with the power

$$P_v = \sin^2\theta_1 \cdot \cos^2\theta_2 B(\theta_3, \cdots, \theta_N) \tag{21}$$

Clearly $\tilde{y}_j(t)$, $\tilde{y}_l(t)$ and $v(t)$ are linearly dependent, and thus, at most, two of them could belong to the optimum set. Also, $\theta_1 \geqq \theta_2$ and therefore

$$P_v \geqq P_j \tag{22}$$

Thus $\tilde{y}_j(t)$ has the smallest power among the three dependent signals and can be excluded from the optimum set, contrary to our assumption.

QED. Lemma 1

LEMMA 2

There exists an optimum solution satisfying

$$\hat{\theta}_k \leqq \frac{\pi}{4} \qquad k = 2, \cdots, N \tag{23}$$

Proof

Let $\hat{\theta}_i$, $i = 1, \cdots, N$ be an optimum set of modulation angles and suppose we have, for some k, $\hat{\theta}_k > \pi/4$.

Define

$$\tilde{\theta}_k = \frac{\pi}{2} - \hat{\theta}_k < \frac{\pi}{4} \tag{24}$$

Since $\sin^2\tilde{\theta}_k = \cos^2\hat{\theta}_k$ and $\cos^2\tilde{\theta}_k = \sin^2\hat{\theta}_k$ we could modify our data set to satisfy

$$\begin{aligned} \tilde{P}_j &= \sin^2\tilde{\theta}_k A_j = \cos^2\hat{\theta}_k A_j = \hat{P}_j \\ &\text{and} \\ \tilde{P}_j &= \cos^2\tilde{\theta}_k A_j = \sin^2\hat{\theta}_k A_j = \hat{P}_j \end{aligned} \tag{25}$$

Repeating for all $\hat{\theta}_j > \pi/4$ we obtain a new set of N signals, having the same data power (and thus optimum) and satisfying $\tilde{\theta}_k \leqq \pi/4$, $k = 2, \cdots, N$.

QED. Lemma 2

Consider the optimum solution satisfying (by Lemma 2):

$$\left.\begin{aligned} &\hat{\theta}_1 > \pi/4 \\ &\tilde{\theta}_k \leqq \pi/4, \qquad k = 2, \cdots, N \end{aligned}\right\} \tag{26}$$

by Lemma 1,

$$\hat{y}_k(t) = x_1(t)\, z_k(t), \qquad k = 1, \cdots, N$$

For $k \geqq 2$ we have $\cos^2\hat{\theta}_k \geqq \sin^2\hat{\theta}_k$, and arguing as in Case I, we can show that each $x_k(t)$ appears once and only once, thus

$$\hat{y}_k(t) = x_1(t)\, x_k(t), \qquad k = 2, \cdots, N \tag{27}$$

The Nth signal is $\hat{y}_1(t) = x_1(t)$, which has the highest power, since

$$\begin{aligned} P_1 &= \sin^2\theta_1 \prod_{k=2}^{N} \cos^2\theta_k \\ &> \sin^2\theta_1 \cdot \sin^2\theta_j \prod_{\substack{k=2 \\ k\neq j}}^{N} \cos^2\theta_k \\ &= P_j \end{aligned} \tag{28}$$

QED.

4. Conclusion

Given the design conditions, that is the power ratios in the data channels, and the minimum required RF power, it was shown that the optimum modulation scheme is either the Interplex or the conventional method. Which of the two is superior depends on the design conditions; also, other configurations might achieve the same performance. However, in no case can another configuration out-perform both Interplex and conventional system. Therefore, to find the best performance achievable by an N-channel phase-shift-keyed/phase-modulated digital coherent system, the investigation can be narrowed to these two schemes.

Reference

1. Lindsey, W. C., "Design of Block-Coded Communication Systems," *IEEE Trans. Commun. Tech.*, Vol. COM-15, No. 4, August 1967.

D. The Critical Problem and Coding Theory,

N. White

1. Introduction

In this paper we shall investigate in detail the interesting fact that the "critical problem" of combinatorial geometry contains as a special case the problem of determining the largest dimension possible for a linear code over $GF(q)$ of fixed length and minimum distance. In *Subsection 2* we introduce the reader to the critical problem and its relationship to coding theory, and present a sketch of Crapo and Rota's solution of the critical problem in terms of the characteristic polynomial of a certain geometric lattice. In the remainder of the paper we restrict attention to $q = 2$; i.e., binary codes. In *Subsection 3* we show that the critical problem for distance three (single-error correcting) codes has a nice solution in terms of combinatorial geometry. And while this problem has been solved before, our proof is instructive and shows how the highly developed techniques of combinatorial geometry are beginning to cast light on coding theory. In *Subsection 4* we give techniques which are useful for computing the characteristic polynomials for distance-d codes, $d \geqq 4$, and show how the calculations for a code of block length n can in fact be performed in a space of $n - 3$ dimensions. These results take on especially satisfying form for $d = 5$; we show the equivalence of the $d = 5$ problem to a certain generalized coloring problem. One hopes that the $d = 5$ coding problem can be solved in this manner. At present it is unsolved, and the techniques of conventional combinatorial coding theory seem exhausted on this subject. Finally, in *Subsection 5* we obtain a recursion for the characteristic polynomial for $d = 4$ codes. The paper concludes with a short table of the known characteristic polynomials, and should provide useful information from which future theorems may be induced.

2. The Critical Problem

Let V_q^n be a vector space of dimension n over $GF(q)$, the finite field of order q. If $x \varepsilon V_q^n$, we denote

$$x = (x^{(1)}, x^{(2)}, \cdots, x^{(n)})$$

Let S be any set of nonzero vectors in V_q^n, thought of as *error patterns*. We define an *optimal linear code with respect to* S to be any subspace $U \subseteq V_q^n$ of maximum dimension such that $S \cap U = \phi$. For any S we denote some such subspace by $U(S)$, even though U in general is not unique. If $k = k(S) = \dim(U(S))$, then $U(S)$ is an (n, k) linear (block) code on q symbols such that no two of its code words differ by a vector in S. Thus, the code can be used to detect error patterns from S.

Several examples we will use are given below; other classes, such as those arising from burst errors, are also of interest.

$$B_w^n = \{x \varepsilon V_q^n \mid x \neq 0, \quad \text{Hamming weight } (x) \leqq w\}$$

$$D_w^n = \{x \varepsilon V_q^n \mid \text{Hamming weight } (x) = w\} = B_w^n - B_{w-1}^n$$

$$E_w^n = \{x \varepsilon V_q^n \mid x \neq 0, \quad \text{Hamming weight } (x) \text{ is even and} \leqq w\}$$

Thus $U(B_w^n)$ is a linear code of maximum rate with given block length n and minimum weight $w + 1$. We note that

if $\zeta: V_2^n \to V_2^{n+1}$ is the operation of adding a parity check bit to a binary code word, then $\zeta(V_2^n) \cong E_{n+1}^{n+1}$, and $\zeta(B_w^n) \cong E_w^{n+1}$ if w is even.

Any subspace of dimension ℓ in V_q^n is the intersection of $n - \ell$ hyperplanes. Thus the problem of finding $k(S)$ is equivalent to finding the minimum number of hyperplanes $H_1, H_2, \cdots, H_c$ such that $H_1 \cap H_2 \cap \cdots \cap H_c \cap S = \phi$, by taking $U(S) = H_1 \cap H_2 \cap \cdots \cap H_c$. Crapo and Rota (Ref. 1) call this problem the *critical problem* of combinatorial geometry, with the solution $c = c(S) = n - k(S)$ called the *critical exponent*. We now sketch Crapo and Rota's solution to the critical problem in terms of the characteristic polynomial of S.

We begin with the partially ordered set $L(S)$ of all subspaces of V_q^n which are spanned by elements of S, ordered by inclusion. By intersecting these subspaces with S, we have equivalently $L(S) = \{A \subseteq S \mid A = \bar{A}\}$, again ordered by inclusion, where $\bar{A} = (\text{span } A) \cap S$ is the *closure* of $A \subseteq S$. $L(S)$ is a finite geometric lattice with *bottom element* $0 = \phi$ and *top element* $1 = S$. However, we will restrict our notation to the case $L(S)$, so that the reader will require no previous knowledge of geometric lattices or the equivalent concept of combinatorial geometries (Ref. 1), although the concepts are more general.

The *rank function* $r: L(S) \to \mathbb{Z}$, (the integers) is given, for $x \in L(S)$, by $r(x) = \text{dimension } (x)$, considering x as a sub-set of V_q^n. It is elementary to observe that $r(x) = k$ if and only if x has a *basis* of cardinality k, where a basis is an independent subset of maximum cardinality.

On $L(S)$, or more generally, on any locally finite partially ordered set, we have the *Möbius function* (Ref. 2) $\mu_{L(S)}(x, y): L(S) \times L(S) \to \mathbb{Z}$, satisfying

(1) $\mu_{L(S)}(x, y) = 0$ unless $x \leq y$ in $L(S)$

(2) $\mu_{L(S)}(x, x) = 1$ for all $x \in L(S)$

(3) $\sum_{y: x \leq y \leq z} \mu_{L(S)}(x, y) = 0$ if $x < z$ in $L(S)$

The *characteristic polynomial* of $L(S)$, $p(\lambda; L(S)) = p(\lambda; S)$ is defined by

$$p(\lambda; L(S)) = \sum_{x \in L(S)} \mu_{L(S)}(0, x) \lambda^{r(1) - r(x)}$$

Theorem 1. *(Ref. 1, p. 16.4.) The number of ordered sequences $L_1, L_2, \cdots, L_d$ of d linear functionals*

$$L_i: V_q^n \to GF(q)$$

not necessarily distinct, such that

$$\left(\bigcap_{i=1}^{d} \ker L_i \right) \cap S = \phi$$

is the number $q^{d(n - r(S))}\, p(q^d; S)$. Thus $p(q^d; S) = 0$ for $d = 0, 1, 2, \cdots, c(S) - 1$, and $p(q^d; S) > 0$ for $d \geq c(S)$.

It is interesting to note that we get information about the largest subspace of $V_q^n - S$ by studying the internal geometric structure of S, as expressed in $L(S)$.

We now examine the critical problem for the case $q = 2$, $S = B_w^n$. We let $a_{ij \cdots k} \in V_2^n$ be the binary vector

$$a_{ij \cdots k}^{(\ell)} = 1 \text{ if } \ell = i, j, \cdots, k$$
$$= 0 \text{ otherwise}$$

3. The Critical Problem for B_2^n, $q = 2$

The following example is not new, but it motivates the study of the critical problem.

Lemma 2 *$L(B_2^n) \cong P_{n+1}$, the lattice of partitions of a set of $n + 1$ elements, where P_{n+1} is ordered by $\sigma_1 \leq \sigma_2$ if σ_1 is a refinement of σ_2.*

Proof. $B_2^n \cong E_2^{n+1} = D_2^{n+1}$. Let $X = \{x_1, x_2, \cdots, x_{n+1}\}$ and let A be a closed subset of D_2^{n+1}. We define the blocks of a partition $\sigma(A)$ on X to be the equivalence classes of the relation $x_i \sim x_j$ if and only if $a_{ij} \in A$ or $i = j$. It is clear that $\sim$ is an equivalence relation; in particular, if $x_i \sim x_j$ and $x_j \sim x_k$, then either $i = j$ or $j = k$, hence $x_i \sim x_k$; or else $a_{ij}, a_{jk} \in A$, hence $a_{ik} = a_{ij} + a_{jk} \in A$; and again $x_i \sim x_k$. Thus $\sigma: L(B_2^n) \to P_{n+1}$.

On the other hand, given $\pi \in P_{n+1}$, let

$$\sigma^{-1}(\pi) = \{a_{ij} \in S_2^n \mid x_i \approx x_j, i \neq j\}$$

where $\approx$ is the equivalence relation whose classes are the blocks of π. Let $a_{kl} \in$ span $(\sigma^{-1}(\pi))$. Since we are working over $GF(2)$, a minimal subset of $\sigma^{-1}(\pi)$ whose span contains a_{kl} must be a simple circuit with a_{kl} removed, $a_{lm}, a_{mn}, \cdots, a_{hk}$. Thus $x_l \approx x_k$ and $a_{kl} \in \sigma^{-1}(\pi)$, and $\sigma: L(B_2^n) \cong P_{n+1}$.

QED.

But $P_{n+1} \cong LC(C_{n+1})$, the lattice of contractions of the complete graph C_{n+1} on $n + 1$ vertices, by contracting all edges between vertices corresponding to elements of the same block of a partition. Here $LC(G)$ is ordered so that G itself is the bottom element, for any graph G.

The reader may have noticed that the weight-two vectors D_2^{n+1} correspond in the obvious way to the edges of C_{n+1} under the above isomorphisms. Similarly, any graph G on $n+1$ vertices corresponds to some subset $S_G \subseteq D_2^{n+1}$, with $L(S_G) \cong LC(G)$. Furthermore, $p(\lambda; S_G) = 1/\lambda\, C(\lambda; G)$ where $C(\lambda; G)$ is the chromatic polynomial of G (e.g., Ref. 3, pp. 234–235).

Thus for $G = C_{n+1}$

$$p(\lambda; B_2^n) = (\lambda - 1)(\lambda - 2) \cdots (\lambda - n)$$

Hence by Theorem 1

$$k(B_2^n) = n - c(B_2^n) = n - [\log_2 n] - 1$$

It is interesting to note that the solution of the critical problem for all S_G such that G is a planar graph is equivalent to the famous four-color problem, since G is four-colorable if and only if $C(S_G) \leq 2$.

4. The Critical Problem for B_w^n, $w \geq 3$, $q = 2$

By the following proposition, it suffices to consider only the case that w is even.

Proposition 3. *$k(B_{2l}^n) = k(B_{2l+1}^{n+1})$ for l, $n \geq 0$, hence $C(B_{2l}^n) = C(B_{2l+1}^{n+1}) - 1$.*

Proof. The parity check operator

$$\zeta : V_2^n \underset{\rightarrow}{\cong} E_{n+1}^{n+1}$$

as defined in *Subsection 1*, induces $B_{2l}^n = E_{2l}^{n+1}$ and $U(B_{2l}^n) \cong U'$ for U' a subspace of E_{n+1}^{n+1} missing E_{2l}^{n+1},

$$r(U') = r[U(B_{2l}^n)]$$

Thus U' has minimum weight $\geq 2l + 2$, and $k(B_{2l+1}^{n+1}) \geq k(B_{2l}^n)$.

Since $B_{2l+1}^{n+1} \neq \phi$,

$$k = r[U(B_{2l+1}^{n+1})] < n + 1$$

hence $U(B_{2l+1}^{n+1})$ can be expressed as the row space of a k by $(n+1)$ matrix M which has dependent columns. Hence we can delete one column to get a matrix M' with rank $M' =$ rank M. Let $U' =$ the row space of M', $U' \subseteq V_2^n$. Then U' has minimum weight $\geq 2l + 1$, since $U(B_{2l+1}^{n+1})$ has minimum weight $\geq 2l + 2$, and

$$r(U') = r[U(B_{2l+1}^{n+1})]$$

Thus $k(B_{2l}^n) \geq k(B_{2l+1}^{n+1})$, establishing equality. QED.

Lemma 4. *$c(B_w^n)$ and $k(B_w^n)$ are monotone increasing in n, and*

$$c(B_w^{n+1}) - c(B_w^n) = 1 - (k(B_w^{n+1}) - k(B_w^n)) = 0 \text{ or } 1$$

Proof. Since

$$c(B_w^n) + k(B_w^n) = n$$

it suffices to show

$$k(B_w^n) \leq k(B_w^{n+1}) \leq k(B_w^n) + 1$$

By the isomorphism

$$V_2^n \cong H_{n+1} \subseteq V_2^{n+1}$$

where

$$H_{n+1} = \{x \varepsilon V_2^{n+1} \mid x^{(n+1)} = 0\}$$

$k(B_w^{n+1}) \geq k(B_w^n)$ is trivial.

It suffices to show that

$$\text{rank}\,(U(B_w^{n+1}) \cap H_{n+1}) \geq k(B_w^{n+1}) - 1$$

since

$$k(B_w^n) \geq \text{rank}\,(U(B_w^{n+1}) \cap H_{n+1})$$

Let $U(B_w^{n+1})$ have the basis

$$B = \{b_1, b_2, \cdots, b_k\}, \qquad k = k(B_w^{n+1})$$

Now we assume that some

$$b_i^{(n+1)} \neq 0, \qquad 1 \leq i \leq k$$

for otherwise $U(B_w^{n+1}) \subseteq H_{n+1}$ and we are done. Suppose

$$B' = \{b'_j \varepsilon U(B_w^{n+1}) \cap H_{n+1} \mid b'_j = b_j + b_j^{(n+1)} b_i, \quad j \neq i, \quad 1 \leq j \leq k\}$$

is dependent. Then for some

$$J \subseteq \{1, 2, \cdots, i-1, i+1, \cdots, k\}$$

we have

$$\sum_{j \varepsilon J} b'_j = 0$$

Then

$$\sum_{j \varepsilon J} b_j + \delta b_i = 0$$

for some $\delta \varepsilon GF(2)$, contradicting the independence of B. Thus B' is independent, and

$$\text{rank}\,(U(B_w^{n+1}) \cap H_{n+1}) \geqq |B'| = k - 1. \qquad \text{QED.}$$

If L is any lattice, and $x, y \varepsilon L, x \leqq y$, then we define the *interval lattice*

$$[x, y]_L = \{z \varepsilon L \,|\, x \leqq z \leqq y\}$$

Lemma 5. *If $x \varepsilon L(S)$, $S \subseteq V_q^n$, $r(x) = k$, then $[x, 1]_{L(S)}$ is also a geometric lattice. Indeed, $[x, 1]_{L(S)} \cong L(R)$ for some $R \subseteq V_q^{n-k}$.*

Proof. Let x have a basis $\{b_1, b_2, \cdots, b_k\}$, and let $x_i = \overline{\{b_{k-j+1}, b_{k-j+2}, \cdots, b_k\}}$, for $j = 1, 2, \cdots, k$. Then

$$[x, 1]_{L(S)} = [x_k, 1] \left([x_{k-1}, 1] \left(\cdot \cdot \cdot \left([x_2, 1] \left([x_1, 1]_{L(S)}\right) \cdots\right)\right)\right)$$

because $0, x_1, x_2, \cdots, x_k = x$ is simply a totally ordered chain between 0 and x in $L(S)$, and subintervals of intervals of L are obviously intervals of L. Each interval in the expression above has height one less than its subscript-lattice, thus it suffices to prove the lemma for the case $r(x) = 1$. Then $x = \{\bar{a}\}$ for some $a \varepsilon S, a \neq 0$.

For some $i \leqq n$, $a^{(i)} \neq 0$. Let $\rho: V_q^n \to V_q^{n-1}$ by

$$\rho(x) = x - \frac{x^{(i)}}{a^{(i)}} a$$

where we identify V_q^{n-1} with

$$\{y \varepsilon V_q^n \,|\, y^{(i)} = 0\}$$

Let $R = \rho(S) \subseteq V_q^{n-1}$. Let $Z \varepsilon [a, 1]_{L(S)}$ and let

$$y \varepsilon (\text{span}\,\rho(Z)) \cap R$$

Then

$$y = \sum_{x_j \varepsilon \rho(Z)} \alpha_j x_j = \rho(y')$$

for some $y' \varepsilon S, \alpha_j \varepsilon GF(q)$. Thus

$$y' = \sum_j \alpha_j x_j' + \delta a$$

for

$$x_j' \varepsilon \rho^{-1}(x_j)$$

and $\delta \varepsilon GF(q)$. But $a \varepsilon Z$. Thus $y' \varepsilon$ (span Z) $\cap$ S, and since Z is closed, $y' \varepsilon Z$. Thus $y \varepsilon \rho(Z)$, proving that $\rho(Z)$ is closed $\subseteq R$. Hence ρ induces a function also denoted ρ; $\rho: [x, 1]_{L(S)} \to L(R)$.

On the other hand, if

$$Y \varepsilon L(R), \rho^{-1}(Y) = (\text{span}\,(Y, a)) \cap S$$

hence $\rho^{-1}(Y) \varepsilon [x, 1]_{L(S)}$ since $\rho^{-1}(Y)$ is obviously closed. Thus $\rho: [x, 1]_{L(S)} \cong L(R)$. QED.

We now use a result of Crapo (Ref. 4, p. 606), specialized to geometric lattices of the type $L(S)$. We note that $T \subseteq S \subseteq V_q^n$ implies $L(T) \subseteq L(S)$ canonically.

Proposition 6. *If $T \subseteq S \subseteq V_q^n$, then*

$$\lambda^{r(S)-r(T)} p(\lambda; T) = \sum_{x \varepsilon M(S,T)} p(\lambda; [x, 1]_{L(S)})$$

where

$$M(S, T) = \{x \varepsilon L(S) \,|\, y \varepsilon L(T), y \leqq x \text{ in } L(S) \text{ implies } y = 0\}$$

We again consider exclusively characteristic 2. Let

$$F_w^n = E_w^n - D_2^n = \left\{x \varepsilon V_2^n \,|\, \text{weight}\,(x) = 4, 6, 8, \cdots, 2\left[\frac{w}{2}\right]\right\}$$

Proposition 7. *For $n \geqq 5$, $h \geqq 2$,*

$$p(\lambda; F_{2h}^n) = p(\lambda; B_{2h}^{n-1}) + \binom{n}{2} p(\lambda; B_{2h}^{n-2}) + \binom{n}{3} p(\lambda; B_{2h}^{n-3})$$

Proof. In Proposition 6, let $T = F_{2h}^n \subseteq S = E_{2h}^n$. No $x \varepsilon M(S, T)$ may contain a_{ij} and a_{kl} for any distinct i, j, k, ℓ, for then $a_{ijkl} = a_{ij} + a_{kl} \varepsilon x$. Therefore, $M(S, T) = \{0;$ the $\binom{n}{2}$ points $a_{ij} \varepsilon S - T = D_2^n;$ and the $\binom{n}{3}$ lines $\{a_{ij}, a_{jk}, a_{ik}\}$ for distinct $i, j, k\}$.

By Lemma 5, $[a_{ij}, 1] \overset{\rho}{\cong} L(R)$, where

$$R = \{x + x^{(i)} a_{ij} \,|\, x \varepsilon S\} = \rho(S)$$

Under the usual identification,

$$V_2^{n-1} = \{x \varepsilon V_2^n \,|\, x^{(i)} = 0\}$$

ρ is the identity on V_2^{n-1}, hence $\rho(E_{2h}^{n-1}) = E_{2h}^{n-1}$. On the remainder

$$\{x \varepsilon E_{2h}^n \mid x^{(i)} = 1\}$$

weight $(\rho(X)) \leqq$ weight (x). Since ρ maintains parity, we see $R = E_{2h}^{n-1}$. Thus

$$[a_{ij}, 1]_{L(S)} \cong L(E_{2h}^{n-1}) \cong L(B_{2h}^{n-2})$$

The line $\{a_{ij}, a_{ik}, a_{jk}\} = \ell_{ijk}$ has a basis $\{a_{ij}, a_{jk}\}$. Thus by Lemma 5 and the preceding case,

$$[\ell_{ijk}, 1]_{L(S)} \cong L(E_{2h}^{n-2}) \cong L(B_{2h}^{n-3})$$

For $n \geqq 5$, $h \geqq 2$, $r(F_{2h}^n) = r(E_{2h}^n)$, so we have

$$p(\lambda; F_{2h}^n) = p(\lambda; B_{2h}^{n-1}) + \binom{n}{2} p(\lambda; B_{2h}^{n-2}) + \binom{n}{3} p(\lambda; B_{2h}^{n-3})$$

QED.

Corollary 8. $c(F_{2h}^n) = c(B_{2h}^{n-3})$, hence $k(F_{2h}^n) = 3 + k(B_{2h}^{n-3})$.

Proof. Let $\sigma: V_2^{n-3} \underset{\rightarrow}{\subseteq} V_2^n$ by $[\sigma(x)]^{(i)} = 0$ if $i \geqq n-2$. Then, in V_2^n,

$$U' = \text{span}\,(\sigma(U(B_{2h}^{n-3})), a_{n-2}, a_{n-1}, a_n)$$

contains only vectors of weights $\geqq 2h + 1$ or $\leqq 3$ in V_2^n. Thus

$$U' \cap F_{2h}^n = \phi, k(F_{2h}^n) \geqq 3 + k(B_{2h}^{n-3})$$

and

$$c(F_{2h}^n) \leqq c(B_{2h}^{n-3})$$

On the other hand, if $c = c(B_{2h}^{n-3})$, let

$$Q_c(\lambda) = (\lambda - 1)(\lambda - 2)(\lambda - 4) \cdots (\lambda - 2^{c-1})$$

By Theorem 1

$$Q_c(\lambda) \mid p(\lambda; B_{2h}^{n-3})$$

hence by Lemma 3

$$Q_c(\lambda) \mid p(\lambda; B_{2h}^{n-2}) \text{ and } p(\lambda; B_{2h}^{n-1})$$

Hence

$$Q_c(\lambda) \mid p(\lambda; F_{2h}^n)$$

and

$$c(F_{2h}^n) \geqq c(B_{2h}^{n-3}) \qquad \text{QED.}$$

It is difficult to use Proposition 6 to excise vectors of weight greater than two, since vectors of weight greater than those in S will usually be introduced into the geometries $[x, 1]_{L(S)}$.

Perhaps the most interesting case of the preceding proposition and corollary is $2h = 4$, since $F_4^n = S_4^n$. This set of vectors is transitive under permutation of the coordinates, and with such symmetry one might expect a solution similar to that of $D_2^n \cong B_2^{n-1}$ in *Subsection 3*. Such a solution would immediately give the optimum rate of a 2-error correcting code. The following observation, which generalizes the equivalence between the critical problem for D_2^n and the coloring problem for C_n, may be a step in that direction.

Let L be a linear functional

$$\sum_{i=1}^{n} \alpha_i x^{(i)}, \qquad \alpha_i \varepsilon GF(2)$$

The *Boolean function corresponding to* L is the function

$$\alpha_L = (\alpha_1, \alpha_2, \cdots, \alpha_n): N = \{1, 2, \cdots, n\} \to GF(2)$$

where $\alpha_L(i) = \alpha_i$. The linear functionals on V_2^n are in one-to-one correspondence with the Boolean functions $\alpha: N \to GF(2)$. Let

$$Y_L = \{i \varepsilon N : \alpha_i = 1\}$$

and for $x \varepsilon S$, let

$$E_x = \{i \varepsilon N : x^{(i)} = 1\}$$

Then

$2^{d(n-r(S))}\, p(2^d; S)$ = the number of ordered sequences of d linear functionals $L_1, \cdots, L_d$ such that

$$\left(\bigcap_{i=1}^{d} \ker Li \right) \cap S = \phi$$

= the number of ordered sequences of d subsets (including ϕ) $Y_1, Y_2, \cdots, Y_d$ of N such that for each $x \varepsilon S$

$|Y_j \cap E_x|$ is odd for some j, $1 \leqq j \leqq d$

We consider the binary relation $R \subseteq A_1 x A_2$ where $A_1 = 2^N = \{Y \subseteq N\}$ and $A_2 = \{E_x \subseteq N : x \varepsilon S\}$, and YRE_x, for $Y \varepsilon A_1$, $E_x \varepsilon A_2$, if and only if $|Y \cap E_x|$ is odd. Then $c(S) =$ minimum $\{|A'| : A' \subseteq A_1$, and every $E_x \varepsilon A_2$ is related to some $Y \varepsilon A'\}$. For $S = D_4^n$, $A_2 = \{X \subseteq N : |X| = 4\}$, and the binary relation should be easier to handle than the corresponding relation for $S = B_4^n$.

5. Recursion for $p(\lambda; B_3^n)$

We know $k(B_3^n)$ from *Subsection* 3 by applying Proposition 3, but that does not give us the characteristic polynomial. Using Proposition 6, we obtain a recursion for $p(\lambda; B_3^n)$ involving $p(\lambda; D_3^n)$.

Let $T = B_3^n - D_2^n \subseteq B_3^n = S$ in Proposition 6. The rank k elements of $M(S, T)$ are the closures of independent sets of k weight-two vectors, which are partitions of an n-element set into $n - k$ parts, as in *Subsection* 3. $[a_{ij}, 1]_{L(B_3^n)} \cong L(B_3^{n-1})$ as in Proposition 7, since $\rho(B_3^{n-1}) = B_3^{n-1}$ and ρ does not increase weights. Thus if $r(x) = k$, $x \varepsilon M(S, T)$, $[x, 1]_{L(B_3^n)} \cong L(B_3^{n-k})$. Thus

$$p(\lambda; T) = \sum_{k=0}^{n-1} S(n, n-k)\, p(\lambda; B_3^{n-k})$$

where

$S(n, m) =$ the Stirling number of the second kind

$=$ the number of partitions of an n-element set into m parts

We re-apply Proposition 6 to the case $T = D_3^n$,

$$S = B_3^n - D_2^n = D_3^n \cup D_1^n$$

$M(S, T) =$ {the 0-, 1- and 2-element subsets of D_1^n}. $[a_i, 1]_{L(S)} \cong L(B_3^{n-1})$ because in the notation of Lemma 5,

$$\rho(D_3^{n-1} \cup D_1^{n-1}) = D_3^{n-1} \cup D_1^{n-1}$$

since ρ fixes x if $x^{(i)} = 0$. But for $a_{ijk} \varepsilon D_3^n \subset S$, $\rho(a_{ijk}) = a_{jk}$.

Therefore $\rho(S) = B_3^{n-1}$. Similarly

$$[\{a_i, a_j\}, 1]_{L(S)} \cong L(B_3^{n-2})$$

Thus we have

$$\begin{aligned} p(\lambda; D_3^n) &= p(\lambda; B_3^n - D_2^n) + np(\lambda; B_3^{n-1}) + \binom{n}{2} p(\lambda; B_3^{n-2}) \\ &= np(\lambda; B_3^{n-1}) + \binom{n}{2} p(\lambda; B_3^{n-2}) + \sum_{k=1}^{n} S(n, k)\, p(\lambda; B_3^k) \\ &= np(\lambda; B_3^{n-1}) + \binom{n}{2} p(\lambda; B_3^{n-2}) \\ &\quad + \text{the } n\text{th coefficient of} \\ &\quad \sum_{k=1}^{n} \frac{p(\lambda; B_3^k)\, x^k}{(1-x)(1-2x)\cdots(1-kx)} \end{aligned}$$

The term for $k = n$ is just $p(\lambda; B_3^n)$, which may be transposed with $p(\lambda; D_3^n)$ to give

$$\begin{aligned} p(\lambda; B_3^n) &= p(\lambda; D_3^n) - np(\lambda; B_3^{n-1}) - \binom{n}{2} p(\lambda; B_3^{n-2}) \\ &\quad - \text{the } n\text{th coefficient of} \\ &\quad \sum_{k=1}^{n-1} \frac{p(\lambda; B_3^k)\, x^k}{(1-x)(1-2x)\cdots(1-kx)} \end{aligned}$$

Table 1. Values of $\tilde{p}$

n	$c(B_4^n)$	$k(B_4^n)$	$\tilde{p}(\lambda; B_4^n)$	$c(D_4^n)$	$k(D_4^n)$	$\tilde{p}(\lambda; D_4^n)$
4	4	0	1	1	3	λ^2
5	4	1	$(\lambda - 15)$	2	3	$(\lambda - 1 \pm i)$
6	5	1	$(\lambda - 25)$	3	3	$(\lambda - 4 \pm 2i)$
7	6	1	$(\lambda - 35)$	4	3	$(\lambda - 10 \pm 4.5i)$
8	6	2	$(\lambda - 49.5 \pm 8.3i)$	4	4	$(\lambda - 15.5)(\lambda - 19.75 \pm 9.9i)$
9	7	2	$(\lambda - 64 \pm 16.7i)$	5	4	$(\lambda - 30.3)(\lambda - 32.3 \pm 18.8i)$
10	7	3	$(\lambda - 95)(\lambda - 81.5 \pm 31.4i)$	6	4	$(\lambda - 51.6)(\lambda - 47.7 \pm 32.2i)$
11	7	4	$(\lambda - 83.3 \pm 44.6i)(\lambda - 133.7 \pm 44.98i)$	6	5	$(\lambda - 68.0 \pm 1.6i)(\lambda - 65.5 \pm 50.0i)$
12	8	4		7	5	$(\lambda - 101.75 \pm 34.1i)(\lambda - 82.25 \pm 67.7i)$
n	$c(B_3^n)$	$k(B_3^n)$	$\tilde{p}(\lambda; B_3^n)$	$c(D_3^n)$	$k(D_3^n)$	$\tilde{p}(\lambda; D_3^n)$
3	3	0	1	1	2	1
4	3	1	$(\lambda - 7)$	1	3	$(\lambda - 1)^3$
5	4	1	$(\lambda - 10)$	1	4	$(\lambda - 2.08 \pm .4i)(\lambda - 2.4 \pm 2.35i)$
6	4	2	$(\lambda - 13 \pm 2.45i)$	1	5	$(\lambda - 1.99)(\lambda - 387 \pm 1.05i)(\lambda - 4.6 \pm 5.3i)$
7	4	3	$(\lambda - 14.8)(\lambda - 16.6 \pm 5.0i)$	1	6	$(\lambda - 2.0002)(\lambda - 3.95)(\lambda - 6.8 \pm 2.5i)(\lambda - 7.2 \pm 9.2i)$

6. Values for $\tilde{p}$

Table 1 gives a set of values for

$$\tilde{p}(\lambda; S) = \frac{p(\lambda; S)}{(\lambda-1)(\lambda-2)(\lambda-4)\cdots(\lambda-2^{\sigma(S)-1})}$$

for $S = B_4^n, D_4^n, B_3^n, D_3^n$. We denote $(\lambda - a + bi)(\lambda - a - bi)$ by $(\lambda - a \pm bi)$.

References

1. Crapo, H. H., and Rota, G.-C., *Combinatorial Geometries*, preliminary edition, M.I.T. Press, Cambridge, Mass., 1970.
2. Rota, G.-C., "On the Foundations of Combinatorial Theory" I, *Z. Wahrschein. und Verw. Gebiete*, Vol. 2, pp. 340–368, 1964.
3. Ore, O., "Theory of Graphs," *AMS Colloquium Publications*, Vol. 38, 1962.
4. Crapo, H. H., "Mobius Inversion in Lattices," *Archiv der Math.*, Vol. 19, pp. 595–607, 1968.